DeepPolicyTracker: Tracking Changes In Environmental Policy In The Brazilian Federal Official Gazette With Deep Learning

Flávio Nakasato Cação [1] Anna Helena Reali Costa [1] Natalie Unterstell [2] Liuca Yonaha [2] Taciana Stec [2] Fábio Ishisaki [2]

Abstract

Even though most of its energy generation comes from renewable sources, Brazil is one of the largest emitters of greenhouse gases in the world, due to intense farming and deforestation of biomes such as the Amazon Rainforest, whose preservation is essential for compliance with the Paris Agreement. Still, regardless of lobbies or prevailing political orientation, all government legal actions are published daily in the Federal Official Gazette. However, with hundreds of decrees issued every day by the authorities, it is absolutely burdensome to manually analyze all these processes and find out which ones can pose serious environmental hazards. In this paper, we propose the DeepPolicyTracker, a promising deep learning model that uses a state-of-the-art pretrained natural language model to classify government acts and track changes in the environmental policies. We also provide the used dataset annotated by domain experts and show some results already obtained. In the future, this system should serve to scale up the high-quality tracking of all oficial documents with a minimum of human supervision and contribute to increasing society's awareness of every government action.

1. Introduction

Brazil has one of the largest reserves of biodiversity in the world, such as the Amazon rainforest, Cerrado and Atlantic forest. The preservation of these biomes is essential for the country to be able to fulfill the objectives of the Paris Agreement (Rochedo et al., 2018), since 78% of greenhouse gas emissions in Brazil come from land use and cover change (West et al., 2019). In 2020, while global emissions fell as a result of the coronavirus pandemic, in Brazil they grew substantially driven by deforestation and farming (Spring, 2020); the Amazon Rainforest deforestation rate was the greatest of the decade (Silva Junior et al., 2021). At the same time, the country is an agribusiness powerhouse, with 26.6% of the GDP related to it (CEPEA-ESALQ, 2021) and is governed by a president who did not hide his intentions to weaken current environmental policies (Nature, 2018). In this complex environment, it is extremely important, and it is a high leverage strategy, to be able to track the acts issued by the government, in order to alert and empower civil society with qualified and clear information (Rolnick et al., 2019). However, as noted in (Grimmer & Stewart, 2013), this is an arduous task for manual work alone: hundreds of highly technical documents are issued every day by the Congress and the Executive branches.

Therefore, these issues represent an opportunity for the most recent models that automate the natural language processing (NLP) tasks. Pretrained language models have started to become popular in recent years and have set new quality standards in virtually all NLP tasks such as classification, translating and question answering. They have millions or billions of parameters and are built upon the Transformer architectures (Vaswani et al., 2017). These models are pretrained in a self-supervised way over huge databases, like the entire Wikipedia, and then they can be used to solve other language problems, with fine-tunings in smaller domain-specific datasets. The result obtained in general is much superior when compared to a model trained solely on the smaller dataset (Raffel et al., 2020).

In this work, we contribute with the *DeepPolicyTracker* model, an approach to classify federal government acts using a state-of-the-art NLP technique, called BERT (Devlin et al., 2019), which is a bidirectional encoder architecture. In order to fine-tune this architecture to our specific scope, a preliminary dataset was prepared where thousands of documents were initially pre-classified and filtered by a rule-based robot and then reviewed and enriched under the supervision of domain experts. It is noteworthy that the system formed only by the rule-based robot followed by a layer of human supervision today feeds one of the

[1]Polytechnic School, University of Sao Paulo, Sao Paulo, Brazil [2]Política por Inteiro, Sao Paulo, Brazil. Correspondence to: Flávio Nakasato Cação <flavio.cacao@usp.br>.

Tackling Climate Change with Machine Learning Workshop at ICML 2021.

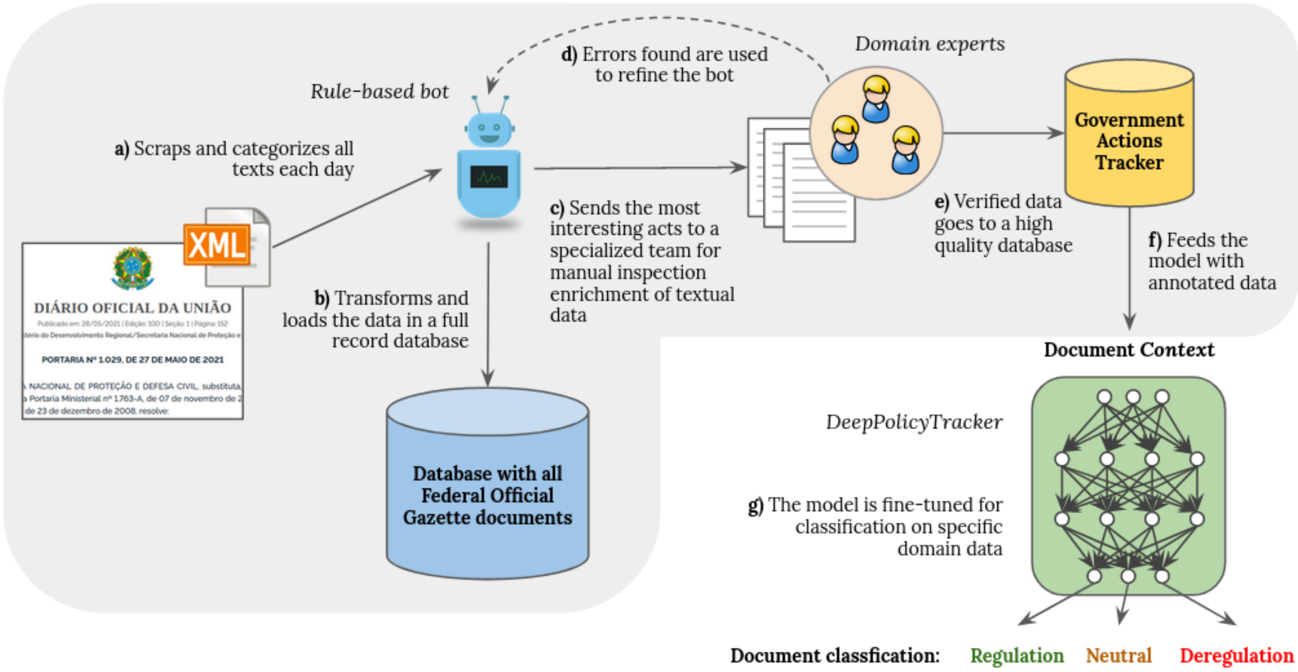

Figure 1. The operational flow of the data pipeline. A rule-based robot scrapes and pre-classifies all official documents released every day (step *a*), and loads them into a general database (step *b*). The most relevant acts related to the environment are sent to a team of domain experts (step *c*), who manually reviews the robot's classifications and enriches the database with new information. Errors found by annotators are regularly used to improve the rule-based robot (step *d*). This filtered and verified subset of data is loaded into the *Government Actions Tracker* database (step *e*), which is used to fine-tune the BERT classification model. This model receives a *Context* variable for each document and is fine-tuned to classifies it as a Regulation/Deregulation or Regulation/Deregulation/Neutral action (step *g*). The system over the gray area represents the Política Por Inteiro's system currently deployed in production.

largest newspapers in the country with daily monitoring of acts by the Brazilian government that may have negative consequences for the preservation of the country's native forests and wildlife[1]. Thanks to this, it was possible to identify massive repeals of protection laws moved by the Federal Government in 2020, with the potential increase in deforestation[2]. We hope that our new proposal with the deep learning BERT module can give greater efficiency and effectiveness to this task.

In the next sections, we will cover the construction strategy of the dataset used for training the BERT network, as well as the configurations used in its training.

[1]The Environmental Policy Monitor can be accessed here: https://arte.folha.uol.com.br/ambiente/monitor-politica-ambiental

[2]A newspaper article reporting this can be found here: https://www1.folha.uol.com.br/ambiente/2020/07/governo-acelerou-canetadas-sobre-meio-ambiente-durante-a-pandemia.shtml

2. Methods

2.1. Data Preparation

Every morning, a robot scrapes all documents published in the Federal Official Gazette[3] and pre-classifies them under "Themes", based on rules defined by domain experts and refined over the years. These rules are based mainly on keywords and more complex expressions to include or exclude a document from a given theme. So far, there are 23 possible themes like *Climate Change*, *Amazon Region* and *Environmental Disasters*. All official document data is transformed and loaded into a database. This is illustrated in Figure 1.

The most relevant documents filtered by the robot are also sent to a separate file, where, every day, two specialists jointly review them and annotated an *Action*, a *Circumstance* and a *Classification* fields for each record, besides some more useful metadata. An *Action* refers to the legal action defined by the document, while a *Circumstance* usually

[3]The official documents of the federal government, originally in PDF, are also published in a machine readable format – in this case, XML. These are the files processed by the robot.

carries more details about the action taken. Both are, for the most part, only extracted from the original document with minimal adjustments, and concatenated into a new variable *Context* we created to feed the BERT model. Regarding the *Classification* field, domain experts defined 12 classes, described below:

- **Regulation:** Action that seeks to institute a rule or norm by the public administration, giving guidelines and producing guidance to economic agents;
- **Deregulation:** Action that seeks to revoke and/or reverse a previously established regulation, change its understanding or orientation;
- **Institutional reform:** Change in structure, skills and institutional arrangement related to public policy;
- **Response:** Action that aims to respond to a significant external event, such as a natural disaster or a major accident;
- **Flexibilization:** Alteration, temporary or not, of deadlines or conditions for compliance with environmental rules, norms and legislation;
- **Neutral:** Action with no significant impact when considered in isolation, but cataloging assessed as necessary because it addresses topics on relevant agendas or with indications of becoming relevant in the medium and long terms;
- **Retreat:** Action that seeks to revoke, replace or modify previously established regulations, due to political or popular pressure;
- **Law consolidation:** Result of regulatory review, with no impact on content;
- **Revocation:** Batch revisions or acts associated with the full revision process;
- **Privatization:** Action that seeks the alienation of business rights under the competence of the Union; the transfer, to the private sector, of the execution of public services operated by the Union; or the transfer or grant of rights over movable and immovable property of the Union;
- **Legislation:** Action that seeks to agree a new law before society, giving guidelines and providing guidance to economic agents;
- **Planning:** Action that does not institute regulatory processes per se, but discloses documents and guiding strategies, such as management plans, creation of committees and working groups, approval of programs and policies that have not yet been defined, among others.

Misclassifications found by the annotators are also used regularly to refine the rule-based robot.

After the human supervision stage, the verified and enriched data are sent to a separate database, the Government Actions Tracker database [4]. Our main contribution is the fine-tune of a BERTimbau Base model, a pretrained BERT model in Portuguese with 12 layers and 110 million parameters (Souza et al., 2020), on the Government Actions Tracker database to predict the class of a document in the *Classification* field given the *Context* variable created. Since the current process *requires* the evaluation of human experts, a more effective classification system could eliminate the need for human supervision in the vast majority of cases, allowing their efforts to be redirected to new challenges, and dramatically scaling the model's tracking capability.

2.2. Experiment Description

Due to the small size of the database, the training with the original 12 classes proved to be very unstable. Thus, we regrouped the previous classes of the *Classification* variable into two training settings:

- **With three major classes:**
  - **Regulation:** Regulation, Planning and Response
  - **Neutral:** Neutral, Retreat and Legislation
  - **Deregulation:** Privatization, Deregulation, Flexibilization, Institutional reform, Law consolidation and Revocation
- **With only two major classes:**
  - **Regulation:** Regulation, Planning, Response, Neutral, Retreat and Legislation
  - **Deregulation:** Privatization, Deregulation, Flexibilization, Institutional reform, Law consolidation and Revocation

We obtained better and more stable results with a batch size of 8, a maximum input sequence length of 200 and a learning rate of 5e-5. Data were shuffled and splitted into training (80%), validation (10%) and test (10%), and the test set was isolated to prevent data leakage. Balancing classes by assigning different weights also showed no significant improvement.

3. Results

The Table 1 summarizes the Matthews Correlation Coefficient (MCC) and the Accuracy (Acc) results obtained for

[4]Due to the curatorial process involved, this database (which is made available at https://github.com/nakasato/deeppolicytracker) today has around one thousand instances, while the complete database has almost 700 thousand – both increase every day.

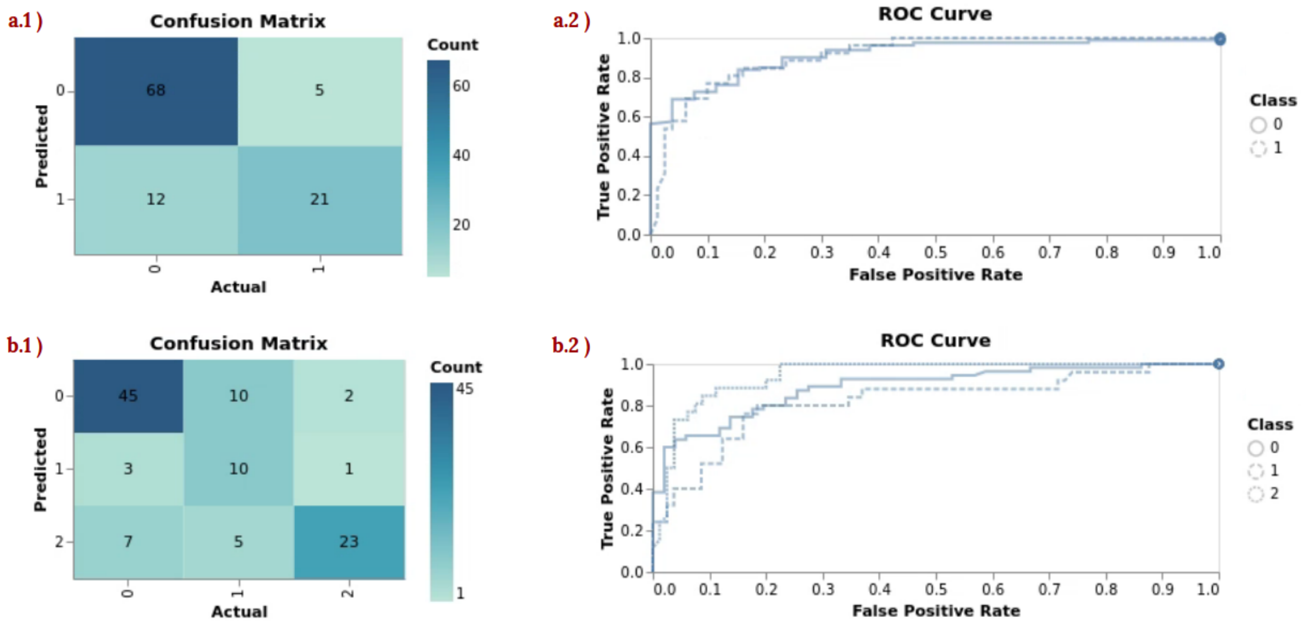

Figure 2. Confusion matrices and ROC curves generated by the models a) with 2 classes (0: *Regulation*, 1: *Deregulation*) and b) with 3 classes (0: *Regulation*, 1: *Neutral*, 2: *Deregulation*).

Table 1. Summary of the Matthews Correlation Coefficient (MCC) and Accuracy (Acc) results on the test set for the two types of classification, with 2 and 3 classes.

Target	MCC	Acc	Proportion of each class.
Classification (2 classes)	0.61	0.84	Regulation (68.7%), Deregulation (31.3%)
Classification (3 classes)	0.57	0.74	Regulation (49.3%), Neutral (19.4%), Deregulation (31.3%)

the two types of rearrangements, one with the original 12 *Classification* classes aggregated in only 2 groups, and the other with them aggregated in 3 groups. Figure 2 shows the confusion matrices and the ROC curves generated in each case. All the presented results were obtained on the test set. Considering the proportions of each class in each experiment, one can see that the *DeepPolicyTracker* model is promising, despite the challenges of dealing with an extra small database for the current standards of the state-of-the-art pretrained models and also despite of the imbalanced classes. These results are particularly notable considering that it is a completely end-to-end system.

4. Conclusion and Future Work

We present the *DeepPolicyTracker*, a model in progress of an end-to-end neural system based on state-of-the-art NLP which aims to track and classify potentially harmful changes in environmental policies directly from texts in official documents. Despite the challenges contained in an extra small base, documents full of jargon and with imbalanced classes, the system shows promising results. Monitoring each act published by the government in order to inform civil society is an extremely challenging task and there is still no single practical solution to solve it. Whether a rule-based system working jointly with domain experts already deliver immeasurable value when it comes to policy monitoring, it is also paramount that the latest NLP technologies are considered to increase the scalability and performance of these systems. Hence, among the future works are the expansion of the annotated datasets, as well as the improvement of the *DeepPolicyTracker* model so it maintains the quality and stability of the classification for a greater number of classes; to this end, we also intend to test simpler and lighter machine learning models.

5. Acknowledgments

This work is the result of an academic partnership between the Polytechnic School of the University of Sao Paulo and Política Por Inteiro. Without the data curation efforts of Política Por Inteiro experts, building this model would not have been possible. Also, this research was supported by Itaú Unibanco S.A. Findings and conclusions expressed here are the responsibility of the authors.

References

CEPEA-ESALQ. PIB do Agronegócio Brasileiro - Centro de Estudos Avançados em Economia Aplicada - CEPEA-Esalq/USP. 2021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, volume 1, pp. 4171–4186, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics. ISBN 9781950737130. doi: 10.18653/v1/N19-1423. URL http://aclweb.org/anthology/N19-1423.

Grimmer, J. and Stewart, B. M. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013. ISSN 14764989. doi: 10.1093/pan/mps028.

Nature. Brazil's sustainability needs social sciences. *Nature Sustainability*, 1(11):607–607, nov 2018. ISSN 2398-9629. doi: 10.1038/s41893-018-0183-0. URL http://www.nature.com/articles/s41893-018-0183-0.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Rochedo, P. R. R., Soares-Filho, B., Schaeffer, R., Viola, E., Szklo, A., Lucena, A. F. P., Koberle, A., Davis, J. L., Rajão, R., and Rathmann, R. The threat of political bargaining to climate mitigation in Brazil. *Nature Climate Change*, 8(8):695–698, aug 2018. ISSN 1758-678X. doi: 10.1038/s41558-018-0213-y. URL http://www.nature.com/articles/s41558-018-0213-y.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Karthik Mukkavilli, S., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., Creutzig, F., Chayes, J., and Bengio, Y. Tackling climate change with machine learning. *arXiv*, 2019. ISSN 23318422.

Silva Junior, C. H., Pessôa, A. C., Carvalho, N. S., Reis, J. B., Anderson, L. O., and Aragão, L. E. The Brazilian Amazon deforestation rate in 2020 is the greatest of the decade. *Nature Ecology and Evolution*, 5(2):144–145, 2021. ISSN 2397334X. doi: 10.1038/s41559-020-01368-x. URL http://dx.doi.org/10.1038/s41559-020-01368-x.

Souza, F., Nogueira, R., and Lotufo, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. pp. 403–417. 2020. doi: 10.1007/978-3-030-61377-8_28. URL http://link.springer.com/10.1007/978-3-030-61377-8_28.

Spring, J. Deforestation boosts Brazil greenhouse gas emissions as global emissions fall — Reuters, 2020. URL https://www.reuters.com/article/us-brazil-environment-emissions-idUSKBN22X2AA.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):5999–6009, 2017. ISSN 10495258.

West, T. A. P., Börner, J., and Fearnside, P. M. Climatic Benefits From the 2006–2017 Avoided Deforestation in Amazonian Brazil. *Frontiers in Forests and Global Change*, 2(September):1–11, 2019. ISSN 2624-893X. doi: 10.3389/ffgc.2019.00052.