Text clustering applied to unbalanced data in legal contexts

Lucas José Gonçalves Freitas
Brazilian Supreme Federal Court - STF/ Brasília, Distrito Federal - Brasil
Brasilia University - UnB/ Brasília, Distrito Federal - Brasil
lucas.freitas@stf.jus.br

Abstract

The Supreme Federal Court (STF), the highest judicial instance in Brazil, generates an immense amount of data organized in text format, including decisions, petitions, injunctions, appeals, and other legal documents, much like lower-level courts. These documents are grouped and classified by specialized employees involved in legal process initiation (case filing), who, in specific cases, utilize technological tools for support. Some cases that reach the STF, for instance, are categorized under one or more Sustainable Development Goals (SDGs) from the United Nations' 2030 Agenda. This categorization aims to facilitate internal and external assessments of the court's performance in addressing the central themes of the Agenda. Given the manual and repetitive nature of this task and its connection to pattern detection, it is feasible to develop machine learning and artificial intelligence-based tools for this purpose. In this study, Natural Language Processing (NLP) models are proposed for process clustering with the goal of augmenting the database concerning certain Sustainable Development Goals (SDGs) with limited recorded occurrences. The clustering or grouping activity, which is highly significant in its own right, can also bring unlabeled entries around processes already categorized by the Court team. This, in turn, enables new labels to be assigned to similar processes. The results obtained demonstrate that augmented sets through clustering can be utilized in supervised learning workflows to assist in case classification, especially in contexts with imbalanced data. This extended abstract shares all bibliographic references with the original dissertation, which is cited here in the references section.

1 Methodology

The objective of this study is to employ clustering methods to bring together labeled and unlabeled texts related to the United Nations' 2030 Agenda Sustainable Development Goals (UN General Assembly, 2015). The aim is to utilize the proximity of labeled processes in strategies for data augmentation based on the propagation of synthetic labels. By the end of the proposed workflow, it is expected that the synthetic labels generated through clustering will ease the challenge of classifying imbalanced labels, a common occurrence in 2030 Agenda Sustainable Development Goals (SDGs) with limited natural entries in the Supreme Federal Court's (STF) procedural classification service.

Text classification algorithms for 2030 Agenda SDGs are applied within the RAFA 2030 (Artificial Networks with a Focus on the 2030 Agenda in Portuguese) initiative, an artificial intelligence tool currently in use at the court. The RAFA 2030 initiative's application (RAFA 2030, 2022), developed using the Shiny package in the R language, includes neural network-based label suggestions and graphical decision support tools. These tools encompass co-occurrence graphs, bigram wordclouds, and specialized searches for laws and legal articles.

In summary, the proposed data augmentation strategy in this study serves the purpose of enhancing the classifiers currently employed at the court by balancing classes with few records, as illustrated in Figure 1.

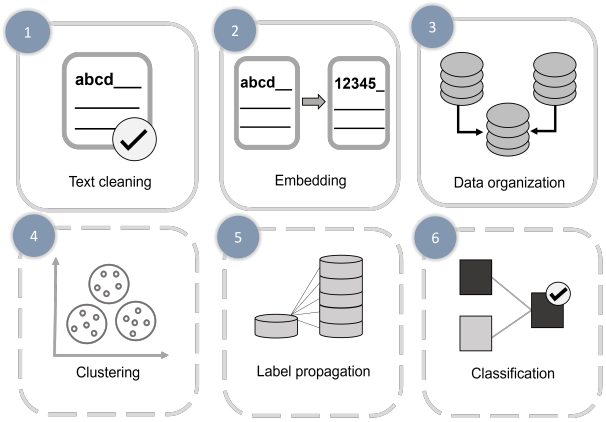

Figure 1: Basic flowchart

The steps denoted by solid lines represent stages applied to the entire dataset, in other words, across all SDGs in study. The steps identified by dashed lines are carried out on a per-SDG basis. The embedding model mentioned in step 2, based on the doc2vec algorithm, has also been employed in another artificial intelligence initiative at the Supreme Federal Court called vitorIA (Supreme Federal Court, 2023). Its primary objective is to group similar texts for subsequent batch processing and the identification of potential topics of general repercussion or repetitive issues in legal cases.

In broad strokes, the proposed workflow initiate with the data cleaning phase, followed by the embedding step. At this point, the dataset comprising processes that have been evaluated and categorized by court employees under the 2030 Agenda SDGs is combined with another dataset, with no original labels. The concept is straightforward: processes without original labels may receive synthetic labels depending on their proximity to originally evaluated processes, rendering the augmented datasets less imbalanced and containing a greater number of examples. Enhanced datasets lead to improved predictions by classification models, which is the final step in the workflow outlined in this study.

Data

The originally labeled dataset comprises approximately 2,000 petitions and rulings from the Supreme Federal Court. Petitions serve as legal documents initiating court cases, while judgments are documents produced by the courts themselves after the initial decision in a case. The assessments of SDGs in this set of documents were carried out by experts within the court.

On the other hand, the unlabeled dataset consists of 40,000 rulings from cases not previously labeled for the 2030 Agenda SDGs. By utilizing the same data processing and embedding mechanism for both datasets, we create a larger dataset comprising approximately 42,000 labeled and unlabeled processes. This dataset, particularly the embedding vectors associated with each text, plays a crucial role in the clustering step.

It's worth noting that while initial petitions hold significance, rulings contain nearly all available information in the legal cases. This is because judgments are rendered after decisions have been made, and all arguments have been presented and evaluated by the judges. Another advantage of using rulings relates to document formatting. Judgments (rulings) are produced within the courts themselves, making them more conducive to PDF reading and processing.

Text cleaning and Embedding

The texts were subjected to a standard natural language processing cleaning process. This included the removal of portuguese and legal stopwords, converting text to lowercase, removing accents and special characters. Additionally, during the text reading and OCR process, non-relevant objects such as headers, file margins, branded symbols, signatures, and other irrelevant graphical elements were eliminated to enhance the comprehension of the texts themselves. To further condense the texts, parts of speech tagging steps can be applied to retain only nouns, adverbs, adjectives, and verbs, using pre-trained portuguese-based dictionaries. This represents an aggressive cleaning approach that has shown excellent performance, especially in lengthy legal documents.

The embedding step was carried out using the doc2vec algorithm, which, despite being created in 2018, remains highly relevant for large texts. This is particularly true for texts that do not perform well with frequency-based embeddings like TF-IDF. Adjusted with smaller windows than the default settings, this embedding model has been utilized in recursive process clustering strategies (ARE, AI, RE procedural classes) within the scope of the Supreme Federal Court, integrated into the vitorIA tool.

Clustering

The document vectors are clustered using the k-means algorithm, and the determination of the number of clusters to be formed is a crucial parameter in the proposed strategy. The choice of a straightforward clustering method aligns with the same rationale behind selecting the doc2vec algorithm for embedding. The primary aim of this research is to establish a baseline assessment of the proposed strategy through the utilization of simple methods, thereby providing the flexibility to incorporate more sophisticated techniques throughout the entire workflow when necessary. The naturally obtained clusters bring together labeled, unlabeled, and not evaluated processes. Synthetic labels are propagated as depicted in Figure 2.

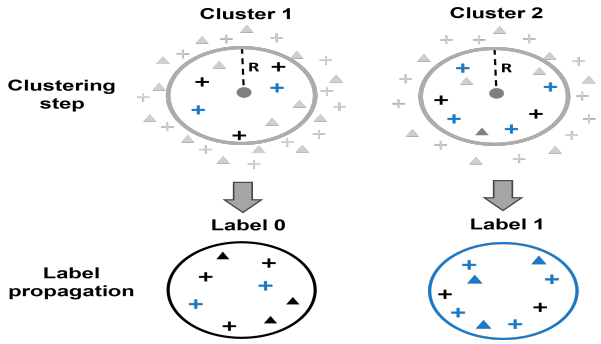

Figure 2: Label propagation strategy

The gray triangles represent not evaluated processes, while the blue crosses denote evaluated processes with labels and the black crosses signify evaluated processes without labels. By establishing intervals for the radius R, it becomes possible to avoid the periphery of clusters, where, theoretically, process vectors from one group exhibit greater similarity to processes from another cluster. Within the set defined by radius R, the proportion of labeled neighbors (threshold) determines the label propagation for all not evaluated processes within the set. This is a straightforward propagation approach but serves the purpose of creating a baseline for the strategy effectively. All parameters (number of clusters, radius R, and threshold) were selected within specified ranges using typical machine learning strategies based on train-validation-test procedures. The augmented datasets remain imbalanced but provide a larger number of examples for the algorithms employed in the classification task. Table 1 displays the distribution of labels before and after augmentation with synthetic labels.

	Original dataset		Augmented dataset	
SDGs	Labels 0	Labels 1	Labels 0	Labels 1
SDG 3	1635	370	3590	590
SDG 4	1877	128	3908	509
SDG 8	1559	446	3453	654
SDG 9	1937	68	3964	548
SDG 10	1635	370	3604	642
SDG 11	1914	91	3953	535
SDG 15	1909	96	3934	529
SDG 16	763	1242	3957	6438
SDG 17	1787	218	3692	4355

Table 1: Label distribution in original and augmented databases

SDG	Clusters	Radius (%)	Threshold (%)
SDG 3	25	10	60
SDG 4	25	10	60
SDG 8	25	10	70
SDG 9	25	10	70
SDG 10	25	10	70
SDG 11	25	10	70
SDG 15	25	10	70
SDG 16	25	25	60
SDG 17	50	25	60

Table 2: Parameter selection in clustering validation

It is possible to observe that some Sustainable Development Goals (SDGs) significantly increased the example base, with records showing up to 5 times more cases with labels. The substantial increase in examples within broader and more generic SDGs is of particular interest to legal actors, as in such cases, categorization can be more complex when carried out through subjective means. The difference in the total number of processes in the augmented dataset for each SDG is a result of the synthetic label propagation strategy itself. The augmented datasets are then employed for training LSTM networks, which are currently in use within the court (as part of the RAFA 2030 initiative). Among the 17 SDGs, those not assessed in this study had very few labeled examples at the time, necessitating a preliminary step to handle small sample sizes.

2 Results

The primary outcomes of the clustering phase entail the selection of optimal parameters for each of the assessed Sustainable Development Goals (SDGs). Table 2 presents the final parameters for each sustainable development objective, obtained during the validation step.

The number of clusters remains constant at 25,

except for SDG 17. Here, 5, 10, 25, 50, or 100 clusters were evaluated. The radii for escaping the cluster limit range from 10% to 25% of the processes closest to the cluster center. Evaluations were conducted on the 5%, 10%, 25%, 50%, and 100% of processes closest to the centroid, with 100% indicating the entire cluster. The label propagation threshold vary between 60% and 70%. Proportions of 50%, 60%, and 70% of neighboring processes with labels were analyzed for label propagation within not evaluated cases of a cluster.

The metrics obtained from the adjustment of LSTM neural networks for the original and augmented datasets are presented in Table 3. It can be observed that there is a expressive improvement in some SDGs with limited natural entries.

	Original dataset		Augmented dataset	
SDG	Accuracy	Sensitivity	Accuracy	Sensitivity
SDG 3	0.83	0.80	0.89	0.82
SDG 4	0.79	0.83	0.84	0.81
SDG 8	0.86	0.81	0.87	0.83
SDG 9	0.81	0.79	0.89	0.87
SDG 10	0.83	0.79	0.85	0.79
SDG 11	0.78	0.75	0.82	0.81
SDG 15	0.72	0.72	0.83	0.83
SDG 16	0.87	0.82	0.91	0.85
SDG 17	0.73	0.75	0.74	0.76

Table 3: LSTM neural net performance - original and augmented datasets

3 Conclusions and future works

This work is connected to two artificial intelligence initiatives of the Supreme Federal Court - RAFA 2030 and vitorIA. RAFA 2030 is based on text classification related to the Sustainable Development Goals (SDGs) of the 2030 Agenda, while vitorIA is focused on text clustering for the identification of repetitive demands in legal cases. Regarding the technique presented, it can be observed that data augmentation flows based on text clustering can serve as a treatment for imbalanced datasets with limited entries for a specific class. The strategy for classifying legal cases according to the SDGs of the 2030 Agenda, currently in use at the court, has shown improvements of up to 17% for certain sustainable development objectives, which is a significant outcome. Further approaches can be explored for the embedding, clustering, and propagation of synthetic labels, as this work represents just the baseline. Future research involves the use of Large Language Models (LLM), as well as graph-based strategies for label propagation.

References

UN General Assembly, Transforming our world: The 2030 Agenda for Sustainable Development, 21 October 2015, available at: https://tinyurl.com/dck7yjpv [29 October 2023]

RAFA 2030 (2022). Redes Artificiais com Foco na Agenda 2030, available at: https://github.com/agenda2030rafa [29 October 2023]

Supreme Federal Court (2023). vitorIA, available at: https://tinyurl.com/2wv7vzz5 [29 October 2023]

Text clustering applied to unbalanced data in legal contexts. Msc dissertation, available at: https://tinyurl.com/mtzyuuay