BBRC: Brazilian Banking Regulation Corpora

Rafael Faria de Azevedo, Thiago Henrique Eduardo Muniz,
Claudio Pimentel, Guilherme José de Assis Foureaux,
Bárbara Caldeira Macedo, Daniel de Lima Vasconcelos
Banco do Brasil S.A
SAUN Quadra 5, Lote B, s/n, 70.040-912, Asa Norte, Brasília/DF, Brasil
rafael.f.azevedo@outlook.com
{rafael.azevedo, thiagobancodobrasil, clpimentel,
guilhermefx, barbaracaldeira, delima}@bb.com.br

Abstract

We present BBRC, a collection of 25 corpus of banking regulatory risk from different departments of Banco do Brasil (BB). These are individual corpus about investments, insurance, human resources, security, technology, treasury, loans, accounting, fraud, credit cards, payment methods, agribusiness, risks, etc. They were annotated in binary form by experts indicating whether each regulatory document contains regulatory risk that may require changes to products, processes, services, and channels of a bank department or not. The corpora in Portuguese contain documents from 26 Brazilian regulatory authorities in the financial sector. In total, there are 61,650 annotated documents, mostly between half and three pages long. The corpora belong to a Natural Language Processing (NLP) application that has been in production since 2020. In this work, we also performed binary classification benchmarks with some of the corpus. Experiments were carried out with different sampling techniques and in one of them we sought to solve an intraclass imbalance problem present in each corpus of the corpora. For the benchmarks, we used the following classifiers: Multinomial Naive Bayes, Random Forest, SVM, XGBoost, and BERTimbau (a version of BERT for Portuguese). The BBRC can be downloaded through a link in the article.

Keywords: banking, corpus, regulatory risk

1. Introduction

Regulation is part of business activities in any industry, including the financial sector. Considering the sheer volume of regulations companies must follow, it is often a manual and onerous process. However, regulation is beneficial to the bank's customers, the market, and even the company itself, as it can even bring profits to the bank (Pasiouras et al. 2009; Aldasoro et al. 2020; Kim et al. 2013). To manage and automate the regulation it must respond to, Banco do Brasil created a tool to manage the daily publications that financial market regulatory authorities make, which can impact the company's activities. This tool is called Radar Regulatório (Regulatory Radar), which acronym is RR. It has been in production since 2020, classifying regulatory documents as relevant or irrelevant to several of its departments individually, with regard to their potential to impact its activities from a regulatory risk perspective. If a document is classified as relevant, it is forwarded to each department that the publication may impact. Therefore, experts in the area can evaluate the document and make the necessary changes to keep the department in compliance with the regulatory authority that published the document.

Radar Regulatório (RR) serves more than 40 company departments. It classifies between 300 and 1,000 regulatory documents daily published by more than 100 regulatory authorities (municipal, state, and federal levels). It is important to follow city and state regulations, in addition to federal ones, as their needs vary depending on its characteristics (Lastra, 2019), especially in a country as large as Brazil. The application works with a hybrid approach with a pipeline composed of Machine Learning (ML) and rules (regular expressions - regex).

When an expert from a department analyzes a document classified by the tool, he points out the correctness or otherwise of the labeling of the document made by RR, and thus annotates the document that will be part of the department's corpus to be used in retraining of its Artificial Intelligence (AI) model. Each department has created its regulatory risk corpus according to this annotation process. The junction of each corpus of many departments gave birth to the Brazilian Banking Regulation Corpora (BBRC), which is **the main contribution of this work**.

The BBRC is a set of 25 regulatory risk corpus (legal/financial data) from different departments of Banco do Brasil. Furthermore, the corpora contain documents from 26 different Brazilian banking/finance regulatory authorities, which can affect the bank's various activities (products, processes, services, and channels) of the bank. The corpora belongs to various departments such as insurance, investments, treasury, accounting, agribusi-

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 150–166

ness, human resources, and others. If on the one hand, BBRC can be useful to explore ML algorithms applied to NLP tasks such as text classification, document analysis, and sentiment analysis, on the other hand, each corpus of BBRC can be used in other areas like sociology, economy and politics, as highlighted in the Section 2 and Section 7.

Our second contribution is a benchmark that compares some models we evaluated for a binary classification task. In one of the benchmarks, we evaluate a strategy to deal with the intraclass imbalance problem present in the entire BBRC corpus (Liu et al., 2021).

The rest of this paper is organized as follows. In Section 2, we introduce related works. In Section 3, we present our corpora. The application that caused the creation of BBRC is presented in Section 4. In Section 5, the experiments performed with some corpus of BBRC and the discussion are presented. Section 6 presents our future work and Section 7 concludes the paper.

2. Related Works

In this section, we mainly present works related to the BBRC, but also some related to Radar Regulatório (Regulatory Radar). We start by presenting some corpus similar to BBRC.

Lima et al. (2020) used machine learning to investigate fraud in the Brazilian public sector. They used a dataset constructed with a source that is also present in the BBRC, which is the Brazilian Official Journal (Diário Oficial da União - DOU). The dataset contains 1,907 annotated risk entries. Sohn et al. (2021) presented the Global Banking Standards QA dataset (GBS-QA), a banking regulation dataset of questions from market players and answers from the Basel Committee on Banking Supervision (BCBS). The corpus was reorganized and verified by financial regulatory experts. In our search, few banking regulatory corpus were found; however, when we searched for financial corpus, the quantity of corpus increased.

Jiang et al. (2020) introduced an automatic financial news dataset annotation through a weakly-supervised hierarchical multilabel classification for the Chinese language. The event FinCausal 2020 Shared Task on Causality Detection in Financial Documents created the FinCausal Corpus (financial news feed) (Mariko et al., 2020). Lefever and Hoste (2016) presented a supervised machine learning approach to economic events detection in newswire text. To do so, a corpus of Dutch financial news articles with ten types of company-specific economic events was annotated. The work of Zmandar et al. (2022) presented CoFiF Plus, a narrative summarization dataset created from financial reports in French. It is made up of 1,703 reports covering a time period of 1995 to 2021. Jabbari et al. (2020) described an ontology of compliance-related concepts and relationships (annotation schema). They also presented an annotated corpus of financial news articles in French for entity recognition and relation extraction. Chen et al. (2021) introduced FINQA, an expert-annotated dataset containing 8,281 financial QA pairs, along with their numerical reasoning processes. It was built based on the earnings reports of S&P 500 companies. The dataset was tested with algorithms such as BERT, RoBERTa and FinBERT. DoRe is a French and dialectal French corpus for NLP analytics in finance, regulation, and investment. It is composed of 2,350 Annual Reports from 336 companies, covering a time frame from 2009 to 2019 (Masson and Paroubek, 2020). In addition to financial corpus, we also found legal corpus in our research, which may be related to regulation or the financial sector.

The area of NLP has long studied legal texts, as well as texts from the health sector and other areas, whether in Portuguese or other languages. Just as in the area of health, law also has a wealth of specific terms (Thompson et al. 2011; Halder et al. 2017; Quochi et al. 2008; Pardelli et al. 2012; Delfino et al. 2018), the same phenomenon occurs in the area of banking regulation.

De Araujo et al. (2020) described Victor, a dataset of digitized legal documents from the Brazilian Supreme Court. The corpus supports two tasks, document classification and theme assignment. LexGLUE is a benchmark dataset to evaluate the performance of NLP methods, especially Large Language Models (LLMs). It is based on seven existing legal NLP datasets in English (Chalkidis et al., 2021). Au et al. (2022) describe E-NER, a publicly available NER dataset. It is based on legal company filings available from the EDGAR dataset of the US Securities and Exchange Commission. Chalkidis et al. (2023) presented LeX-Files, a diverse multinational English legal corpus that includes 11 distinct subcorpora that cover legislation and case law from six primarily English-speaking legal systems (EU, CoE, Canada, US, UK, and India). The work also introduces LegalLAMA, a new probing benchmark suite inspired by LAnguage Model Analysis (LAMA). The LexFiles are compared to the Pile of Law corpus (Henderson et al., 2022), a large legal corpus (256GB dataset).

In addition to the datasets, we also tried to find applications similar to RR. We found the use of multiple classifiers to detect investment rules in long regulatory documents (Mansar and Ferradans, 2018), a Python library for NLP and machine learning for legal and regulatory texts (Bommarito et al., 2018), a semi-supervised text classi-

fication framework for operational risk (Zhou et al., 2020) and approaches with LLMs (Mamakas et al. 2022; Chakravarthy et al. 2020). In the next section, our corpora is presented in detail.

3. The Corpora

This section presents the main subject of this paper, the Brazilian Banking Regulation Corpora (BBRC). The corpora annotation started with the creation of the application Radar Regulatório (Regulatory Radar - RR), which is presented in Section 4. A corpora is a collection of corpus. In Natural Language Processing (NLP), a dataset is called corpus. Each corpus in BBRC belongs to a department of Banco do Brasil and was annotated in a binary way: relevant or irrelevant (in reality, the bank's experts annotate the corpora with scores from 0 to 3, where 0 is irrelevant, 1 is not very relevant, 2 is relevant and 3 is extremely relevant, however, the company decided to share the data in binary format). A corpus, in the context of RR, is a collection of documents from various **regulatory authorities** that could affect the activities of a department. Each document was annotated as belonging to the relevant class or to the irrelevant class. If the document was classified as relevant, it means that the department may have to make changes to its activities to comply with the relevant regulatory document published by the regulatory authority. If the document is classified as irrelevant, no change is needed. All corpora documents are public, as documents published by all regulatory authorities mentioned in this article are valid for all Brazilian banks and financial institutions.

The annotation process for each corpus of each department has been performed by one or more experts from that department since 2020. These experts are responsible for ensuring that the changes demanded in a relevant document are met as required by the regulatory authority. For this reason, the corpora did not pass through an evaluation of the agreement between annotators (inter-annotator agreement). When possible, it is a process that produces a more reliable corpus, where each sample is annotated by different annotators, who follow a rigorous process that helps the annotators make decisions guided by a well-developed guideline (this guideline could not be shared by the company). The agreement between annotators evaluation potentially improves the quality of the corpus and, consequently, the quality of the trained model increases, which can be highly affected by corpus quality, as presented by (Alhamzeh et al. 2022; Artstein 2017; Nowak and Rüger 2010).

However, the quality of the annotation of BBRC is ensured by the consequences that can occur if a mistake is made. Failure can cause expensive fines, restrictions, and sanctions to the bank. No expert wants to live in a situation like this. The correct classification of a regulatory document is the first step that decides whether an action plan must be carried out and executed to change a product, a process, a channel, or a service to keep it in compliance. The BBRC data ranges from June 18, 2020 to August 16, 2023.

The regulatory authorities (regulators) belong to one of these three levels of compliance: federal, state, or municipal (as mentioned by Lima et al. 2020). Examples of regulators are the Brazilian Central Bank (Banco Central do Brasil - BACEN), Brazil's federal revenue (Receita Federal do Brasil - RFB), Legislative Assembly of the State of Mato Grosso (Assembleia Legislativa do Estado do Mato Grosso) and Rio de Janeiro City Council (Câmara Municipal do Rio de Janeiro). Table 1 presents numbers about regulators in BBRC.

From the perspective of all regulatory authorities, BBRC has in total 5,698 unique documents in the relevant class, 20,131 unique documents in the irrelevant class, and 25,829 unique documents considering both classes. To be part of the corpora, each regulator had to have at least five documents classified in the relevant class. Regarding the departments, only those with at least 50 documents of the relevant class were elected to the corpora. The description of all 25 departments (corpus) of the bank in the corpora is given in Table 8, which is in the Appendix A Section at the end of the paper, after the references.

Table 2 presents the description of each column of the corpora. The idea was to offer a wider understanding of the details of the corpora. The corpora[1] is shared with the community in a CSV format file (1.7 GB). Figure 1 presents BBRC data schema. Figures 3, 4, 5, and 6 in the Appendix A present examples of the content of each column of BBRC. Figure 7, also at Appendix A section, presents the text of one sample of the BBRC. Table 3 shows information on the number of samples per class of each corpus in the BBRC.

In total, the corpora has 61,650 document samples, 7,823 in the relevant class, and 53,827 in the other class. The documents are unique in each class and in each corpus, but can be repeated in different corpus. This repetition of documents happens because one document can be relevant or irrelevant for several departments. The most important feature (column) in the BBRC is **text** (as it is an NLP dataset collection).

Table 4 presents the basic statistics of the column text in the relevant class. Character information

[1]Data available at https://huggingface.co/datasets/bancodobrasil/bbrc_brazilian_banking_regulation_corpora

Regulatory authority	Relevant	Irrelevant	Total
National Civil Aviation Agency (ANAC)	14	507	521
Brazilian Association of Financial and Capital Market Entities (ANBIMA)	262	381	643
National Data Protection Authority (ANPD)	7	22	29
National Supplementary Health Agency (ANS)	85	346	431
Legislative Assembly of the State of Mato Grosso	9	66	75
Brazil, Stock Exchange, Counter (B3)	886	1,241	2,127
Brazilian Central Bank (BACEN)	1,796	3,455	5,251
Commodities and Futures Exchange & São Paulo Stock Exchange (BM&F BOVESPA)	13	39	52
National Bank for Economic and Social Development (BNDES)	176	202	378
Rio de Janeiro City Council	11	843	854
Securities Custody and Financial Settlement Center (CETIP) (currently B3)	141	47	188
Federal Accounting Council (CFC)	32	99	131
Interbank Payments Chamber (CIP)	266	640	906
Financial Activities Control Board (COAF)	40	42	82
Accounting Pronouncements Committee (CPC)	18	4	22
Securities and Exchange Commission (CVM)	380	1,026	1,406
Brazilian Official Journal (DOU)	534	7,203	7,737
Brazilian Federation of Banks (FEBRABAN)	8	0	8
National Institute of Information Technology (ITI)	23	56	79
Ministry of Labour	11	44	55
Núclea (previous CIP)	39	62	101
Presidency of the Republic (PR)	176	844	1,020
National Supplementary Pension Superintendence (PREVIC)	22	79	101
Brazil's Federal Revenue (RFB)	370	1,117	1,487
National Treasury Secretariat (STN)	227	1,003	1,230
Private Insurance Superintendency (SUSEP)	152	763	915
Total	**5,698**	**20,131**	**25,829**

Table 1: Column "Relevant" presents unique documents in the relevant class for each regulator. The column "Irrelevant" presents unique documents in the irrelevant class for each regulator. The column "Total" shows the unique documents in the whole corpora for each regulator. The URLs of all regulators are in the Appendix A, Table 9. The name of the regulatory authority was translated, but the acronym was kept in Portuguese.

	Column	Description
1	class	The class of the document is 1 to relevant or 0 to irrelevant
2	department	The department (board, directorate or related company) of Banco do Brasil that uses RR with a corpus
3	entry_date	The date the document was received by RR from the contracted company
4	general_id	The document's unique identifier across the entire corpora
5	normative_identifier	Identifier of the regulatory document given by the regulatory authority
6	publication_date	The date the regulatory document was published
7	regulatory_authority	The regulatory authority (regulator) that published the regulatory document
8	subject	Most regulatory documents usually have a subject, such as a title or summary
9	subject_length	The number of characters in the subject
10	subject_unique_words	The number of unique words in the subject
11	subject_words	The number of words in the subject
12	text	The full text of the regulatory document
13	text_length	The number of characters in the text
14	text_unique_words	The number of unique words in the text
15	text_words	The number of words in the text
16	type	The type of the regulatory document, most regulatory authorities publish several types of documents
17	unique_document_id	Unique identifier of the regulatory document in the corpus (is repeated in different corpus)

Table 2: BBRC columns description.

can give an idea of the length of documents. Assuming that a Microsoft Word page holds around 3,600 characters (Arial 11), the median (middle quartile) of a text in the relevant class is longer than a page. So, 50% of the documents in the relevant class are at least one full page long. Similar statistics occur in the irrelevant class. To count the **words** and **unique words** of each document, a function was used to separate words between blank spaces. All texts were analyzed in their original state (without preprocessing or cleaning), and noise such as URLs, HTML, and email addresses could have caused the incorrect number of words in the text. However, the results still give a fairly precise idea of the size of the document.

The main contribution of this work is BBRC, which is fundamental to Radar Regulatório (Regulatory Radar). This application is presented in the next section.

4. The Application

Before Radar Regulatório (Regulatory Radar - RR), the entire regulatory risk process was done manually, without a formal process, and without standards, where most departments acted in isolation. For example, department A could have one regulatory risk expert who searched and read all regulatory documents published every day to evaluate whether a new norm or a new law could impact the businesses of department A. On the other hand, department B could have a team of three experts that checked once a month for possible

```
Data columns (total 17 columns):
 #   Column                  Dtype
---  ------                  -----
 0   class                   int64
 1   department              object
 2   entry_date              object
 3   general_id              int64
 4   normative_identifier    object
 5   publication_date        object
 6   regulatory_authority    object
 7   subject                 object
 8   subject_length          int64
 9   subject_unique_words    float64
 10  subject_words           float64
 11  text                    object
 12  text_length             int64
 13  text_unique_words       float64
 14  text_words              float64
 15  type                    object
 16  unique_document_id      int64
dtypes: float64(4), int64(5), object(8)
memory usage: 2.7 GB
```

Figure 1: BBRC data schema

impacting regulatory documents that could affect the businesses of department B. One department could have to check 10 regulators' websites, while another department could have to check 50 regulators' websites. This difference occurs because of the characteristics of the business in which the department is involved.

RR was created to solve all these problems. At first, a pure AI (ML) application was thought to solve the issue. After all, AI is widely used in the financial industry (Wall 2018; Zhang et al. 2018). However, the small amount of initial samples and the overlapping classes showed that the use of rules (regex) would also be necessary. So, what worked was a pipeline made up of ML models (Support Vector Machine - SVM) and deterministic rules for a binary classification challenge. Even if the aim of this article is not to present the application, a brief architecture explanation will help with corpora construction understanding. A detailed overview of the application was presented in the article published by de Azevedo et al. (2022).The application is presented in figure 2.

The application pipeline starts at **step 1**, it represents a hired company that collects daily all documents published by all regulators of interest of all departments of the bank (a little more than 100 regulatory authorities have their publications classified daily by the application). In the preprocessing phase (**step 2**), the numbers, special characters, and Portuguese stop words (NLTK) present in the document are removed. All tokens are turned to ASCII version and lowercased. Vectorization is performed by the TF-IDF algorithm, also in Step 2. In steps 3 to 6, the single regulatory document (norm/law) that entered the pipeline will be evaluated in an iterative manner for all models and rules

Department	Relevant	Irrelevant
BB Seguros	184	2,402
BB Asset	702	5,379
CIB	703	1,828
COGER	320	2,366
COGER GESUB	542	2,614
COGER GETRI	137	737
DICRE	88	3,858
DIGOV	253	1,397
DIMEP	581	3,494
DINED	121	9
DIOPE	361	2,333
DIOPE GEFID	403	2,559
DIPES	79	50
DIRAG	345	3,234
DIRIS	184	3,134
DISEM	439	1,566
DITEC	80	30
TESOU	411	3,093
UAC	53	781
UCF	157	129
UCI	193	4,153
UGE	144	363
UNI	94	1,329
UPB/MERCAP	429	2,933
USI	819	4,056
Total	**7,823**	**53,827**

Table 3: Number of samples per class of each department/corpus

Relevant class	Characters	Words	Unique words
Mean	26,221.44	3,786.43	755.22
Standard deviation	98,636.79	13,273.48	1,864.91
Minimum	7	1	1
25% (lower quartile)	1,612	246	150
50% (middle quartile) (median)	3,872	570.5	289
75% (upper quartile)	14,235	2,142	695
Maximum	1,457,062	190,940	31,196

Table 4: Statistics of the **text column** in the **relevant class** in terms of number of characters, number of words, and number of unique words.

of each department registered in the application.

In **step 3** the ML model of a department predicts whether the regulatory document is relevant or irrelevant for the department's business. In **step 4**, there is a rule that has keywords registered for the department that are searched in the text of the document being evaluated. If there is a match, the document is classified as relevant; otherwise, it is

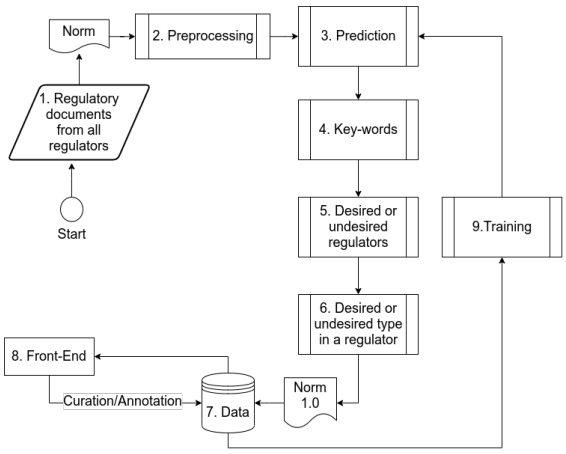

Figure 2: Radar Regulatório architecture. The pipeline each document (norm, law, etc.) passes through for each department.

classified as irrelevant. In the rule of **step 5**, each department can fill 2 lists, one of the desired regulators and another of the undesired ones. If the document evaluated by the rule was published by a regulator in the desired regulators list, the document will be classified as relevant. However, if the document was published by a regulator in the undesired list, it is classified as irrelevant. The same two lists (desired and undesired) in **step 6** are filled only for regulators registered in the desired regulators list of step 5. These lists in step 6 refer to the type of document, as a regulator publishes different types of documents (the type attribute is presented on line 16 of Table 2).

The classification of a previous step can be replaced by the classification done in the current step, except for step 3, which is the first classification. Another point is that step 6 will only activate if the regulator of the document being evaluated is in the desired regulators list of step 5. Once the document is classified, it is saved in the database (**step 7**). The front-end of the application gets all classified documents of each department once a day and presents the relevant ones to the experts of each department (each expert only receives documents of its department). These professionals check the classification of the tool and indicate to the system if it is correct or not (annotation/curation) (**step 8**). From time to time, the ML model of each department is re-trained (**step 9**) with the annotated data stored in the database (step 7).

In summary, Radar Regulatório (Regulatory Radar) classification eases the work of all workers who used to do the same classification process manually. The application prevents errors that could lead to expensive fines and restrictions. In other words, it stops them from having to search for a needle (document) in the haystack once regulators publish far more documents that do not impact the company's businesses (irrelevant documents). The next section presents the benchmarks of the experiments performed with BBRC and the discussion.

5. Experiments and Discussion

This section presents baseline experiments with BBRC using five different algorithms. They are Multinomial Naive Bayes (Kibriya et al., 2005), Random Forest (RF) (Breiman, 2001), Support Vector Machine (SVM) (Cortes and Vapnik 1995; Platt et al. 1999; Chang and Lin 2011), eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016), and BERTimbau (Base and Large) (Souza et al., 2020) (a variation of BERT (Devlin et al., 2018) for Portuguese). The source code for the Machine Learning (ML) experiments is available on GitHub[2]. Figure 8 and Figure 9 show the hyperparameters used in the BERTimbau experiments, which were made with batch_size equals 20, 512 tokens and 5 epochs. The experiments used the same preprocessing (cleaning) described in step 2 of Section 4. The difference is that only documents with at least 50 words (selected before preprocessing and cleaning) were elected to be part of the train and test. In the case of BERTimbau, the stop words were not removed and the characters were not turned into ASCII ones, but the UTF-8 version was kept. For shallow machine learning and deep learning algorithms (BERTimbau), a 60% train and a 40% test split was used. All experiments were carried out using a fixed seed (random_state), and the data was stratified. We conducted two different experiments: a simple binary classification and, a binary classification dealing with the intraclass imbalancing problem. GridSearchCV optimization (Pedregosa et al., 2011) was applied to shallow machine learning algorithms. The hyperparameters of BERTimbau are presented in the Appendix A Section, the ML used algorithms are in the GitHub mentioned above.

All experiments used only samples from the text column of each corpus used. In the first experiment, we made approaches with and without the undersampling technique to evaluate the effect of the inter-class imbalance problem. The first experiment was carried out with 6 corpus of the corpora. They were USI, CIB, BB Asset, DIMEP, COGER GESUB, and DISEM. The second experiment used 3 corpus, which were USI, CIB and, BB Asset.

The **first experiment (experiment 1)** was a binary classification, its results are presented in Table 5. BERTimbau had the best results in all the

[2]Code available at https://github.com/bancodobrasil/bbrc

comparisons in the imbalanced data experiment, which could indicate the superiority of LLMs (deep learning) over shallow machine learning in this scenario. The BERT algorithm and its variations have been successful in many works (Sarkar et al. 2021; Huang et al. 2023; Campiotti et al. 2023). However, when the data was balanced by the undersampling technique, there was no winner algorithm, the results were pretty close, and it showed the impact of undersampling compared to imbalanced data in the results. Furthermore, corpus in Table 5 are sorted in descending order, considering the largest number of samples from the relevant class. The USI corpus is the one with the largest number of samples from the relevant class, and the DISEM corpus is the one with the smallest number of samples from that class (see Table 3). Undersampling was done using the total number of samples from the relevant class in the irrelevant class (819 samples from the USI corpus, 702 samples from the BB Asset corpus, 439 samples from the DISEM corpus, etc. - see Table 3).

In the **second experiment (experiment 2)**, the problem of intraclass imbalance was addressed. This problem exists in all corpus of BBRC. It happens to both classes, relevant and irrelevant. The point is that in the same class there exist more samples from some regulators than samples from other regulators. It happens because some regulators publish far more regulatory documents than others. To carry out the experiment, we chose the four regulators with more documents in the relevant class of each of the three corpus (USI, CIB, and BB Asset). To be part of the evaluation, regulators must have documents in both classes of the corpus. Only regulators with at least 10 documents in the relevant class were chosen. To perform intraclass undersampling, we took the regulator with the fewest documents in the relevant class for each corpus (considering the prerequisites already-mentioned), 26 DOU documents in the USI corpus, for example. Table 6 shows the number of samples available in the corpus chosen for the experiment. Table 7 presents the results of the three corpus evaluated in the second experiment. In this experiment, interclass undersampling was also applied. We observe that shallow machine learning algorithms had better results in corpus such as USI and CIB. However, surprisingly, BERTimbau got the best result in BB Asset, which is exactly the one with fewer samples.

The future work is presented in the next section.

6. Future Work

In future, we intend to expand the corpora with more samples and possibly offer to the scientific community a version of BBRC with scores classification, instead of a binary one. Furthermore,

Corpus	Classifier	F1 score undersampling	F1 score imbalanced
USI	Multin. NB	0.8225	0.7978
	RF	0.8487	0.8097
	SVM	0.8716	0.8456
	XGB	**0.8874**	0.8116
	BERTimbau Base	0.8822	0.8925
	BERTimbau Large	0.8869	**0.8972**
CIB	Multin. NB	0.6691	0.5919
	RF	0.7715	0.595
	SVM	0.7783	0.6485
	XGB	0.7783	0.6758
	BERTimbau Base	**0.782**	**0.7909**
	BERTimbau Large	0.78	0.7634
BB Asset	Multin. NB	0.8532	0.6091
	RF	0.8774	0.4132
	SVM	0.8717	0.6141
	XGB	**0.8884**	0.5465
	BERTimbau Base	0.8829	0.7728
	BERTimbau Large	0.8646	**0.7935**
DIMEP	Multin. NB	**0.8931**	0.538
	RF	0.8742	0.4256
	SVM	0.8919	0.552
	XGB	0.8714	0.508
	BERTimbau Base	0.8749	0.7643
	BERTimbau Large	0.8872	**0.7716**
COGER GESUB	Multin. NB	0.8461	0.5573
	RF	0.8535	0.4368
	SVM	0.8542	0.5155
	XGB	0.8401	0.5014
	BERTimbau Base	**0.855**	0.7155
	BERTimbau Large	0.8117	**0.731**
DISEM	Multin. NB	0.8684	0.6203
	RF	**0.8936**	0.4835
	SVM	0.8557	0.561
	XGB	0.8739	0.5306
	BERTimbau Base	0.8888	0.721
	BERTimbau Large	0.8747	**0.7485**

Table 5: Results of 6 different models trained on 6 corpus of BBRC for the binary classification task.

	Relevant			Irrelevant		
	USI	CIB	BB Asset	USI	CIB	BB Asset
ANBIMA	-	55	221	-	85	399
B3	-	302	147	-	592	1921
BACEN	666	175	-	467	257	
COAF	28	-	-	48	-	-
CVM	-	49	253		368	783
DOU	26	-	-	2,982	-	-
Presidência da República	29	-	-	32	-	-
RFB	-	-	13	-	-	678

Table 6: Quantity of samples per class, per corpus and per regulator used in the second experiment.

Corpus	Samples per class	Unique words	Classifier	F1 score undersampling
USI	104	13,272	**Multin. NB**	**0.6434**
			RF	0.5542
			SVM	0.617
			XGB	0.619
			BERTimbau Base	0.4034
			BERTimbau Large	0.458
CIB	196	21,236	Multin. NB	0.7234
			RF	**0.7283**
			SVM	0.7261
			XGB	0.6499
			BERTimbau Base	0.5895
			BERTimbau Large	0.6494
BB Asset	52	8,472	Multin. NB	0.5
			RF	0.65
			SVM	0.5142
			XGB	0.65
			BERTimbau Base	0.4986
			BERTimbau Large	**0.6666**

Table 7: Results from 6 different classifiers (inductors) trained on 3 BBRC corpus attacking the intraclass problem (binary classification).

we intend to make experiments with BBRC using Generative AI.
The conclusion section ends the paper.

7. Conclusion

We present BBRC, a corpora that brings together several corpus of regulatory risk documents from the Brazilian banking/financial sector. There are 25 corpus from different areas of banking activities such as insurance, agribusiness, human resources, payment methods, security, investments, among others. In this corpora, 26 regulatory authorities in the financial sector are represented. Each corpus was built for binary classification, as they are used at Banco do Brasil in a tool that has been in production since 2020. In total, the corpora has 61,650 documents, all relevant ones were annotated by experts in the area who built each corpus for the needs of their department. We used some of the BBRC corpus to perform binary classification benchmarks with some shallow and deep learning (LLMs) algorithms. We believe that BBRC can help researchers explore ML applied to the regulatory risk and legal field, document analysis, text classification, sentiment analysis (Nopp and Hanbury 2015; Agarwal et al. 2019) and other tasks.
The contribution of the corpora can easily go beyond AI or computer science (Wu and Salomon 2017; Kim et al. 2013; De Masi et al. 2023), as regulatory texts in banking/finance can be used, for example, to assess whether the rent of natural resources is a blessing or a curse for a country that has its economy based on these resources (Tang et al., 2022). The same corpora can be used to analyze the possibility of regulatory lobbying in favor of consolidated financial companies, as a way to prevent new entrants into the sector (Manish and O'Reilly, 2019). The same data can even allow the study of the impact of regulation on national or foreign banks (Wu and Salomon, 2017). Regulatory datasets are essential sources of study to evolve regulations, which are often not prepared for new events outside its context, such as pandemics, climate change (Le Quang and Scialom, 2022), and other crises (Thiemann et al., 2021).
We also hope that the public sharing of BBRC will encourage the sharing of more corpus of banking regulation and other areas, for Portuguese and other languages. Finally, we hope that our data and benchmarks encourage further exploration of better-performing models and techniques. The link to download BBRC is in Section 3.

8. Acknowledgements

The authors thank Banco do Brasil immensely for sharing such an important set of data to promote NLP research for Portuguese. This action demonstrates its commitment and understanding of the industry's active participation in the evolution of science and technology. Special thanks must be given to the Artificial Intelligence and Analytical Unit (**Unidade de Inteligência Artificial e Analítica - UAN**), the Internal Controls Board (**Diretoria de Controles Internos - DICOI**) and the Technology Board (**Diretoria de Tecnologia - DITEC**). We also would like to thank Tiago Nunes Silva and Leonardo Piccaro Rezende for their participation in the production of this paper and Radar Regulatório (Regulatory Radar).

9. Ethical Considerations

Making BBRC available to the scientific community is a rare opportunity for a company in the financial industry to share annotated data, especially in times of the Brazilian General Data Protection Law (Lei Geral de Proteção de Dados - LGPD). This was only possible because all corpora samples are public, since the regulatory documents that make up the corpora were published by regulatory authorities that make public publications, which affect the entire Brazilian financial sector.

10. Bibliographical References

Arvind Agarwal, Aparna Gupta, Arun Kumar, and Srikanth G Tamilselvam. 2019. Learning risk culture of banks using news analytics. *European Journal of Operational Research*, 277(2):770–783.

Iñaki Aldasoro, Leonardo Gambacorta, Paolo Giudici, and Thomas Leach. 2020. Operational and cyber risks in the financial sector.

Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. 2022. It's time to reason: Annotating argumentation structures in financial earnings calls: The finarg dataset. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 163–169.

Ron Artstein. 2017. Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313.

Ting Wai Terence Au, Ingemar J Cox, and Vasileios Lampos. 2022. E-ner–an annotated named entity recognition corpus of legal text. *arXiv preprint arXiv:2212.09306*.

MJ Bommarito, Daniel Martin Katz, and E Detterman. 2018. Lexnlp: Natural language processing and information extraction for legal and regulatory texts. *Research Handbook on Big Data Law*.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Israel Campiotti, Matheus Rodrigues, Yuri Albuquerque, Rafael Azevedo, and Alyson Andrade. 2023. Debertinha: A multistep approach to adapt debertav3 xsmall for brazilian portuguese natural language processing task. *arXiv preprint arXiv:2309.16844*.

Sharanya Chakravarthy, Tushar Kanakagiri, Karthik Radhakrishnan, and Anjana Umapathy. 2020. Domino at fincausal 2020, task 1 and 2: causal extraction system. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 90–94.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. Lexfiles and legallama: Facilitating english multinational legal language model development. *arXiv preprint arXiv:2305.07507*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Pedro Henrique Luz De Araujo, Teófilo Emídio de Campos, Fabricio Ataides Braz, and Nilton Correia da Silva. 2020. Victor: a dataset for brazilian legal documents classification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1449–1458.

Rafael Faria de Azevedo, João Pedro Santos Rodrigues, Mayara Regina da Silva Reis, Claudia Maria Cabral Moro, and Emerson Cabrera Paraiso. 2018. Temporal tagging of noisy clinical texts in brazilian portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 231–241. Springer.

Rafael Faria de Azevedo, Tiago Nunes Silva, Henrique Tibério Brandão Vieira Augusto, Paulo Oliveira Sampaio Reis, Isadora Bastos Chaves, Samara Beatriz Naka de Vasconcellos, Liliany Aparecida dos Anjos Pereira, Mauro Melo de Souza Biccas, André Luiz Monteiro, and Alexandre Rodrigues Duarte. 2022. Banking regulation classification in portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 137–147. Springer.

Sara De Masi, Kose John, Agnieszka Słomka-Gołębiowska, and Piotr Urbanek. 2023. Regulation and post-crisis pay disclosure strategies of banks. *Review of Quantitative Finance and Accounting*, pages 1–33.

Pedro Delfino, Bruno Cuconato, Guilherme Paulino-Passos, Gerson Zaverucha, and Alexandre Rademaker. 2018. Using openwordnet-pt for question answering on legal domain. In *Proceedings of the 9th Global Wordnet Conference*, pages 105–112.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mahmoud El-Haj, Marina Litvak, Nikiforos Pittaras, George Giannakopoulos, et al. 2020. The financial narrative summarisation shared task (fns 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 1–12.

Kishaloy Halder, Lahari Poddar, and Min-Yen Kan. 2017. Modeling temporal progression of emotional status in mental health forum: A recurrent neural net approach. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 127–135.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234.

Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Ali Jabbari, Olivier Sauvage, Hamada Zeine, and Hamza Chergui. 2020. A french corpus and annotation schema for named entity recognition and relation extraction of financial news. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2293–2299.

Hang Jiang, Zhongchen Miao, Yuefeng Lin, Chenyu Wang, Mengjun Ni, Jian Gao, Jidong Lu, and Guangwei Shi. 2020. Financial news annotation by weakly-supervised hierarchical multi-label learning. In *Proceedings of the second workshop on financial technology and natural language processing*, pages 1–7.

Ashraf M Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. 2005. Multinomial naive bayes for text categorization revisited. In *AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17*, pages 488–499. Springer.

Teakdong Kim, Bonwoo Koo, and Minsoo Park. 2013. Role of financial regulation and innovation in the financial crisis. *Journal of Financial stability*, 9(4):662–672.

Rosa M Lastra. 2019. Multilevel governance in banking regulation. *The Palgrave Handbook of European Banking Union Law*, pages 3–17.

Gaëtan Le Quang and Laurence Scialom. 2022. Better safe than sorry: Macroprudential policy, covid 19 and climate change. *International Economics*, 172:403–413.

Els Lefever and Véronique Hoste. 2016. A classification-based approach to economic event detection in dutch news text. In *10th International Conference on Language Resources and Evaluation (LREC)*, pages 330–335. ELRA.

Marcos Lima, Roberta Silva, Felipe Lopes de Souza Mendes, Leonardo R de Carvalho, Aleteia Araujo, and Flavio de Barros Vidal. 2020. Inferring about fraudulent collusion risk on brazilian public works contracts in official texts using a bi-lstm approach. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1580–1588.

Zhining Liu, Pengfei Wei, Zhepei Wei, Boyang Yu, Jing Jiang, Wei Cao, Jiang Bian, and Yi Chang. 2021. Handling inter-class and intra-class imbalance in class-imbalanced learning. *arXiv preprint arXiv:2111.12791*.

Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022. Processing long legal documents with pre-trained transformers: Modding legalbert and longformer. *arXiv preprint arXiv:2211.00974*.

GP Manish and Colin O'Reilly. 2019. Banking regulation, regulatory capture and inequality. *Public Choice*, 180(1-2):145–164.

Youness Mansar and Sira Ferradans. 2018. Sentence classification for investment rules detection. In *Proceedings of the First Workshop on Economics and Natural Language Processing*, pages 44–48.

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. Financial document causality detection shared task (fincausal 2020). *arXiv preprint arXiv:2012.02505*.

Corentin Masson and Patrick Paroubek. 2020. Nlp analytics in finance with dore: a french 250m tokens corpus of corporate annual reports. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2261–2267.

Sérgio Moro, Paulo Cortez, and Paulo Rita. 2016. An automated literature analysis on data mining applications to credit risk assessment. *Artificial Intelligence in Financial Markets: Cutting Edge Applications for Risk Management, Portfolio Optimization and Economics*, pages 161–177.

Clemens Nopp and Allan Hanbury. 2015. Detecting risks in the banking system by sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 591–600.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566.

Gabriella Pardelli, Manuela Sassi, Sara Goggi, and Stefania Biagioni. 2012. From medical language processing to bionlp domain. In *LREC*, pages 2049–2055.

Fotios Pasiouras, Sailesh Tanna, and Constantin Zopounidis. 2009. The impact of banking regulations on banks' cost and profit efficiency: Cross-country evidence. *International review of financial analysis*, 18(5):294–302.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Valeria Quochi, Monica Monachini, Riccardo Del Gratta, and Nicoletta Calzolari. 2008. A lexicon for biology and bioinformatics: the bootstrep experience. In *LREC*. Citeseer.

Rajdeep Sarkar, Atul Kr Ojha, Jay Megaro, John Mariano, Vall Herard, and John Philip McCrae. 2021. Few-shot and zero-shot approaches to legal text classification: A case study in the financial sector. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 102–106.

Kyunghwan Sohn, Sunjae Kwon, and Jaesik Choi. 2021. The global banking standards qa dataset (gbs-qa). In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 19–25.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Chang Tang, Muhammad Irfan, Asif Razzaq, and Vishal Dagar. 2022. Natural resources and financial development: Role of business regulations in testing the resource-curse hypothesis in asean countries. *Resources Policy*, 76:102612.

Matthias Thiemann, Carolina Raquel Melches, and Edin Ibrocevic. 2021. Measuring and mitigating systemic risks: how the forging of new alliances between central bank and academic economists legitimize the transnational macroprudential agenda. *Review of international political economy*, 28(6):1433–1458.

Paul Thompson, John McNaught, Simonetta Montemagni, Nicoletta Calzolari, Riccardo Del Gratta, Vivian Lee, Simone Marchi, Monica Monachini, Piotr Pezik, Valeria Quochi, et al. 2011. The biolexicon: a large-scale terminological resource for biomedical text mining. *BMC bioinformatics*, 12(1):1–29.

Larry D Wall. 2018. Some financial regulatory implications of artificial intelligence. *Journal of Economics and Business*, 100:55–63.

Zheying Wu and Robert Salomon. 2017. Deconstructing the liability of foreignness: Regulatory enforcement actions against foreign banks. *Journal of International Business Studies*, 48:837–861.

Qi Zhang, Jue Wang, Aiguo Lu, Shouyang Wang, and Jian Ma. 2018. An improved smo algorithm for financial credit risk assessment–evidence from china's banking. *Neurocomputing*, 272:314–325.

Fan Zhou, Shengming Zhang, and Yi Yang. 2020. Interpretable operational risk classification with semi-supervised variational autoencoder. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 846–852.

Nadhem Zmandar, Tobias Daudert, Sina Ahmadi, Mahmoud El-Haj, and Paul Rayson. 2022. Cofif plus: A french financial narrative summarisation corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1622–1639.

A. Appendix

	Department	Description
1	BB Seguros	BB Seguridade Participações is a holding company controlled by Banco do Brasil and operates in the insurance business. The group comprises the controlled companies BB Corretora de Seguros e Gestora de Bens and BB Seguros Participações and their subsidiaries.
2	BB Asset	BB ASSET (DTVM): BB Gestão de Recursos - Distribuidora de Títulos e Valores Mobiliários is a company specialized in the management of third-party resources and in the management of investment funds for Banco do Brasil clients
3	CIB	Corporate and Investment Bank (CIB) Board: acts as client, product and channel manager within the Corporate and Large Corporate segments
4	COGER	Accounting Board (COGER): operates within the scope of accounting strategies; standardization, bookkeeping, control and accounting disclosure; accounting statements; tax planning and management; accounting information to the market; and results of availability, integrity, reliability and compliance of accounting information.
5	COGER GESUB	Coger Executive Management that operates within the scope of BB Subsidiaries
6	COGER GETRI	Executive Management at Coger, which operates in the scope of planning, tax management and tax compliance.
7	DICRE	Credit Department (DICRE): Strategic Unit that operates in the management of credit risk, credit portfolios, customer registration, guarantees, parameterization of credit operations and credit limits, as well as the development of solutions for the credit process in the organization.
8	DIGOV	Acts as manager of clients and government products
9	DIMEP	Payment Means and Services Department (DIMEP): acts as product manager and operational support for Business Transactions, within the scope of the cards, vouchers and Instant Payment System (PIX) market
10	DINED	Digital Business Directorate (DINED): acts as strategy, product and channel manager within the scope of new digital business models, covering startups, distribution on BB's digital platforms and digital ecosystems, including Bank as a Service (BaaS)
11	DIOPE	Operations Board (DIOPE): acts as operational support for business transactions, for internal processes and logistics
12	DIOPE GEFID	Executive Management of Fiduciary Services (GEFID), subordinate to DIOPE: operates within the scope of fiduciary services (specialized services, with duties and attributions arising from legislation and market supervisory bodies, which guarantee the security and credibility required by investors)
13	DIPES	Culture and People Management Directorate (DIPES): Strategic unit that operates within the scope of people management, including recruitment and selection, career, training, remuneration and benefits.

14	DIRAG	Agribusiness Board (DIRAG): manages customers, products and operational support for business transactions, within the scope of agribusiness.
15	DIRIS	Risk Management Board (DIRIS): operates in risk management.
16	DISEM	Business Solutions Directorate (DISEM): Strategic unit that acted as product and channel manager within companies (customers) of various sizes (middle, upper middle, and high middle.)
17	DITEC	Technology Directorate (DITEC): Strategic unit that operates within the scope of Information Technology, as well as IT risk management, IT models and projects, etc.
18	TESOU	Global Treasury Unit (TESOU): operates in cash and liquidity management; treasury operations; and financial portfolio management
19	UAC	Service and Channels Unit (UAC): Strategic unit that operates within the scope of channel management, including monitoring the internal and external environment in relation to standards, regulations,and demands for service and relationship channels, standardization of service procedures performed in the customer service and management of banking correspondents, among others.
20	UCF	Cyber and Fraud Prevention Unit (UCF): operates within the scope of document and electronic fraud prevention strategies; digital/cyber security policies, models, methodologies, tools, standards, and instruments; security management in electronic channels; and results of risks and losses incurred in digital/cyber security processes
21	UCI	Fundraising and Investments Unit (UCI): acts as a product and channel manager within the scope of funding and investment products
22	UGE	Related Entities Governance Unit (UGE): Strategic unit that operates within the scope of the governance of related entities and corporate operations.
23	UNI	International Business Unit (UNI): Strategic unit that operates within the scope of product and channel management at the international level.
24	UPB/MERCAP	UCI/MERCAP (UPB/MERCAP): works with fundraising products. It is subordinate to the UCI
25	USI	Institutional Security Unit (USI): operates in managing the security of environments and people, information security, privacy and protection of personal data.

Table 8: Departments (corpus) in the corpora and their description.

	Regulator or Acronym	**URL**
1	ANAC	https://www.gov.br/anac
2	ANBIMA	https://www.anbima.com.br
3	ANPD	https://www.gov.br/anpd
4	ANS	https://www.gov.br/ans
5	Legislative Assembly of the State of Mato Grosso	https://www.al.mt.gov.br/
6	B3	https://www.b3.com.br
7	BACEN/BCB	https://www.bcb.gov.br/

8	BM&F BOVESPA (currently B3)	https://www.b3.com.br/pt_br/regulacao/oficios-e-comunicados/bm-fbovespa/
9	BNDES	https://www.bndes.gov.br/wps/portal/site/home
10	Rio de Janeiro City Council	https://www.camara.rio/
11	CETIP (currently B3)	https://www.b3.com.br
12	CFC	https://cfc.org.br/
13	CIP	https://www2.cip-bancos.org.br/Paginas/Sobre.aspx
14	COAF	https://www.gov.br/coaf
15	CPC	https://www.cpc.org.br/CPC
16	CVM	https://www.gov.br/cvm
17	DOU	https://www.in.gov.br/servicos/diario-oficial-da-uniao
18	FEBRABAN	https://portal.febraban.org.br/
19	ITI	https://www.gov.br/iti
20	Ministry of Labour	https://www.gov.br/trabalho-e-emprego
21	Núclea	https://www.nuclea.com.br/
22	Presidência da República (PR)	https://www.gov.br/planalto
23	PREVIC	https://www.gov.br/previc
24	RFB	https://www.gov.br/receitafederal
25	STN	https://www.gov.br/tesouronacional
26	SUSEP	https://www.gov.br/susep

Table 9: All regulators present at BBRC (mentioned in Table 1).

class	department	entry_date	general_id	normative_identifier	publication_date	regulatory_authority
1	usi	2021-02-10	8290	0106/2021	2021-02-05	BACEN
1	usi	2021-02-12	8291	4888	2021-02-12	BACEN
1	usi	2021-02-12	8292	72	2021-02-12	BACEN
1	usi	2021-03-19	14875	36912	2021-03-19	BACEN
1	usi	2021-03-24	14879	36935	2021-03-24	BACEN
1	usi	2021-03-25	14880	81	2021-03-25	BACEN
1	usi	2021-04-01	24465	93	2021-04-01	BACEN
1	usi	2021-04-14	24518	99	2021-04-14	BACEN
1	usi	2021-04-14	24519	96	2021-04-14	BACEN
1	usi	2021-04-14	24520	86	2021-04-14	BACEN
1	usi	2021-04-14	24521	95	2021-04-14	BACEN
1	usi	2021-04-14	24522	97	2021-04-14	BACEN
1	usi	2021-04-14	24523	98	2021-04-14	BACEN
1	usi	2021-04-22	24530	89	2021-04-22	BACEN

Figure 3: BBRC columns overview (part 1)

subject	subject_length	subject_unique_words	subject_words
Caracterizado fornecimento intempestivo de informações ao Banco Central do Brasil, sobr	468	57.0	88.0
Altera a Resolução nº 4.734, de 27 de junho de 2019, dispondo sobre a realização de novas	218	29.0	35.0
Altera a Circular nº 3.952, de 27 de junho de 2019, dispondo sobre a realização de novas et	245	31.0	38.0
Divulga a realização de leilão de venda conjugado com leilão de compra pós-fixado Selic no	122	15.0	19.0
Divulga comunicado do Grupo de Ação Financeira contra a Lavagem de Dinheiro e o Financi	120	16.0	18.0
Disciplina os processos de autorização relacionados ao funcionamento das instituições de p	224	26.0	32.0
Altera a Instrução Normativa BCB nº 20, que dispõe sobre os limites de valor para as transa	113	21.0	21.0
Divulga a versão 2.0 do Manual de Segurança do Open Banking.	60	10.0	11.0
Divulga a versão 2.0 do Manual de Escopo de Dados e Serviços do Open Banking.	77	13.0	15.0
Altera a Resolução BCB nº 32, de 29 de outubro de 2020, que estabelece os requisitos técr	196	28.0	31.0
Divulga a versão 2.0 do Manual de APIs do Open Banking.	55	10.0	11.0
Divulga a versão 1.0 do Manual de Experiência do Cliente no Open Banking.	73	12.0	13.0
Divulga a versão 2.0 do Manual de Serviços Prestados pela Estrutura Responsável pela Gov	112	15.0	17.0
Altera a Circular nº 3.682, de 4 de novembro de 2013, e seu Regulamento anexo, para disp	449	47.0	72.0

Figure 4: BBRC columns overview (part 2)

text	text_length	text_unique_words	text_words
DEPARTAMENTO DE RESOLUÇÃO E DE AÇÃO SANCIONADORA GERÊNCIA TÉCNICA EM SÃO PAULO DECIS	843	96.0	141.0
O Banco Central do Brasil, na forma do art. 9º da Lei nº 4.595, de 31 de dezembro de 1964, torna público	1808	169.0	330.0
A Diretoria Colegiada do Banco Central do Brasil, em sessão extraordinária realizada em 11 de fevereiro d	3251	243.0	562.0
Aviso de Abertura do Leilão de Câmbio 11/2021 O Departamento das Reservas Internacionais (DEPIN), d	723	80.0	105.0
Comunicamos, com referência ao previsto no art. 39, alínea “g”, inciso I, da Circular nº 3.978, de 23 de jar	1237	92.0	121.0
A Diretoria Colegiada do Banco Central do Brasil, em sessão realizada em 25 de março de 2021, com base	27568	1058.0	4224.0
O Chefe do Departamento de Competição e de Estrutura do Mercado Financeiro (Decem), no uso das atrib	1206	102.0	150.0
Os Chefes do Departamento de Regulação do Sistema Financeiro (Denor) e do Departamento de Tecnolog	22221	1069.0	3214.0
Os Chefes do Departamento de Regulação do Sistema Financeiro (Denor) e do Departamento de Tecnolog	49205	1435.0	7039.0
A Diretoria Colegiada do Banco Central do Brasil, em sessão realizada em 14 de abril de 2021, com base r	7283	395.0	1041.0
Os Chefes do Departamento de Regulação do Sistema Financeiro (Denor) e do Departamento de Tecnolog	19293	1100.0	2785.0
Os Chefes do Departamento de Regulação do Sistema Financeiro (Denor) e do Departamento de Tecnolog	18126	815.0	2657.0
Os Chefes do Departamento de Regulação do Sistema Financeiro (Denor) e do Departamento de Tecnolog	30146	1438.0	4482.0
A Diretoria Colegiada do Banco Central do Brasil, em sessão realizada em 20 de abril de 2021, com base r	9726	532.0	1470.0

Figure 5: BBRC columns overview (part 3)

type	unique_document_id
PROCESSO ADMINISTRATIVO SANCIONADOR	787104
RESOLUÇÃO	787954
RESOLUÇÃO	787955
COMUNICADO	797670
COMUNICADO	798871
RESOLUÇÃO BCB	799306
INSTRUÇÃO NORMATIVA BCB	801512
INSTRUÇÃO NORMATIVA BCB	805258
INSTRUÇÃO NORMATIVA BCB	805257
RESOLUÇÃO BCB	805252
INSTRUÇÃO NORMATIVA BCB	805254
INSTRUÇÃO NORMATIVA BCB	805255
INSTRUÇÃO NORMATIVA BCB	805256
RESOLUÇÃO BCB	808061

Figure 6: BBRC columns overview (part 4)

O Banco Central do Brasil, na forma do art. 9º da Lei nº 4.595, de 31 de dezembro de 1964, torna público que o Conselho Monetário Nacional, em sessão extraordinária realizada em 11 de fevereiro de 2021, com base no disposto nos arts. 4º, incisos VI e VIII, da referida Lei, e 26-A da Lei nº 12.810, de 15 de maio de 2013, R E S O L V E U : Art. 1º A Resolução nº 4.734, de 27 de junho de 2019, passa a vigorar com as seguintes alterações: “Art. 7º-B As instituições financeiras de que trata o art. 7º-A devem estar aptas a cumprir o disposto nesta Resolução a partir da data mencionada no inciso II do art. 11. § 1º A aptidão de que trata o caput será atestada pelo cumprimento, com sucesso, de todas as etapas dos testes homologatórios de integração de que trata o art. 7º-A, conforme cronograma de que trata o inciso I do art. 8º. § 2º O descumprimento de qualquer etapa dos testes homologatórios de que trata o art. 7º-A sujeita as instituições financeiras às sanções e às medidas administrativas previstas na legislação em vigor, bem como, a critério do Banco Central do Brasil, à suspensão provisória da realização das operações de que trata o art. 1º, a partir da data mencionada no inciso II do art. 11. § 3º O Banco Central do Brasil, ao determinar a suspensão de que trata o § 2º, estabelecerá as condições mediante as quais essa suspensão será levantada.” (NR) “Art. 11. ..... I - na data de sua publicação, em relação aos arts. 7º-A, 7º-B, 8º e 9º; e II - em 7 de junho de 2021, em relação aos demais dispositivos.” (NR) Art. 2º Ficam revogados os §§ 4º e 5º do art. 7º-A da Resolução nº 4.734, de 2019. Art. 3º Esta Resolução entra em vigor na data de sua publicação. Roberto de Oliveira Campos Neto Presidente do Banco Central do Brasil\r\n

Figure 7: Text sample of ”unique_document_id” number 787954, annotated as ”relevant” by DICRE and USI. The same sample was annotated as ”irrelevant” by CIB and DIMEP. The document was published by BACEN (Brazilian Central Bank)

```python
def tokenizer_model(self, model_path, df, token):
    tokenizer = BertTokenizer.from_pretrained(BERTIMBAU_TOKENIZER, do_lower_case=True)

    # Tokenizing the training data
    encoded_data_train = tokenizer.batch_encode_plus(
        df[df.data_type=='train'].new_content.tolist(),
        add_special_tokens=True,
        return_attention_mask=True,
        padding=True,
        max_length=token,
        return_tensors='pt',
        truncation=True
    )

    # Tokenizing test data
    encoded_data_val = tokenizer.batch_encode_plus(
        df[df.data_type=='val'].new_content.tolist(),
        add_special_tokens=True,
        return_attention_mask=True,
        padding=True,
        max_length=token,
        return_tensors='pt',
        truncation=True
    )


    input_ids_train = encoded_data_train['input_ids']
    attention_masks_train = encoded_data_train['attention_mask']
    labels_train = torch.tensor(df[df.data_type=='train'].label_num.values)

    input_ids_val = encoded_data_val['input_ids']
    attention_masks_val = encoded_data_val['attention_mask']
    labels_val = torch.tensor(df[df.data_type=='val'].label_num.values)

    dataset_train = TensorDataset(input_ids_train, attention_masks_train, labels_train)
    dataset_val = TensorDataset(input_ids_val, attention_masks_val, labels_val)

    return(dataset_train, dataset_val)
```

Figure 8: BERTimbau hyperparameters (part 1)

```python
def setup_model(self, model_path, dataset_train, dataset_val, epochs):
    model = BertForSequenceClassification.from_pretrained(model_path,
                                                          num_labels=len(LABEL_DICT),
                                                          output_attentions=False,
                                                          output_hidden_states=False)

    device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
    model.to(device)
    batch_size = BATCH_SIZE
    dataloader_train = DataLoader(dataset_train,
                                  sampler=RandomSampler(dataset_train),
                                  batch_size=batch_size)
    dataloader_validation = DataLoader(dataset_val,
                                       sampler=SequentialSampler(dataset_val),
                                       batch_size=batch_size)
    optimizer = AdamW(model.parameters(),
                      lr=1e-5,
                      eps=1e-8)

    scheduler = get_linear_schedule_with_warmup(optimizer,
                                                num_warmup_steps=100,
                                                num_training_steps=len(dataloader_train)*epochs)
    return(model, dataloader_train, dataloader_validation, epochs, scheduler, optimizer, device)
```

Figure 9: BERTimbau hyperparameters (part 2)