Data transparency analysis

Nelson G. C. Guimaraes

Abstract— The project is intended to provide both an in-depth exploration of the data made available by the Brazilian government about access to public information as well as create a service where anyone can check the likelihood of their requests for information being denied.

1 CONTEXT

Calls for an increase in data transparency have been gaining ground over the last few years. Governments and private companies have to be prepared to deal with such societal demands. In order to respond to such demands the government of Brazil has created an initiative that allows for any citizen to request access to public data on the official portal (*Acesso a Information*, n.d.). All the requests along with their response is made available and anyone can download it. All requests are properly labeled according to what kind of data was requested, the status of such request (granted or denied) and the reasons for denying its access.

The initiative came under criticism after it denied a great number of requests for allegedly non justifiable reasons. It would be of great value to have more understanding about the reasons for a given request being denied or not and providing a system where anyone can test if their request is likely to be denied or not before submitting it.

2 PROJECT OVERVIEW

This project provides two main results: 1) an in depth analysis of the texts from the requests and responses and a new taxonomy to classify them, which is independent from the official classification. 2) An API where anyone can send their request text and get the likelihood of it being denied by the government.

3 DATA

The data for this project can be accessed in the same portal (*Acesso a Information*, n.d.) and is available in both csv and xml. It needs to be downloaded manually as there is no API at the moment. The project will base all its deliverables on the data from 2015 to 2022 (19/02/2022), using the files called “Pedidos” or requests in portuguese language, which is the language for all the content of this file. The author will translate the metadata when necessary in order to make this result available to a wider audience.

3.1 EDA

This section is dedicated to exploring the data

3.1.1 Metadata

The request file contains 21 columns, they are detailed in Table 1.

Table 1— Metadata with english explanation

Column name	Data type	Description
'IdPedido'	Integer	Unique identifier of the request
'ProtocoloPedido'	String (17)	Protocol number of the request
'Esfera'	String(30)	Scope of request (eg. Federal, local)
'OrgaoDestinatario'	String (250)	Recipient agency
'Situacao'	String(200)	Request’s status
'DataRegistro'	Date: DD/MM/YYYY HH:MM:SS	Date of registration of request
'ResumoSolicitacao'	String (255)	Request summary
'DetalhamentoSolicitacao'	String (2048)	Request details
'PrazoAtendimento'	Date: DD/MM/YYYY HH:MM:SS	Date limit for responding to the request
'FoiPrrorrogado'	String (3): "Sim"/"Não"	Binary variable that informs if the limit date for responding to the

		request was extended
'FoiReencaminhado'	String (3): "Sim"/"Não"	Binary variable informing if the request was forwarded or not
'FormaResposta'	String (200)	Form of response requested by the solicitor
'OrigemSolicitacao'	String (50) "Balcão SIC" / "Internet"	Informs the origin of the request's medium
'IdSolicitante'	Integer	Unique identifier of requestor
'AssuntoPedido'	String (200)	Subject of the request attributed by the agency
'SubAssuntoPedido'	String (200)	Subject subcategory attributed by the agency
'Tag'	String (1024)	Tags are attributed to the request in order to make classifications that are not in the predefined Subject/Subcategory groups
'DataResposta'	Date: DD/MM/YYYY HH:MM:SS	Date of the response
'Resposta'	String (8000)	Response
'Decisao'	String (100)	Type of response given to the request
'EspecificacaoDecisao'	String (200)	Subtype of response given to the request

3.1.2 Trends

We can see that there are a total of 628.272 requests considering the whole period, and have been increasing over the years up to the end of 2020, where there is a sharp decline, as Figure 1 demonstrates.

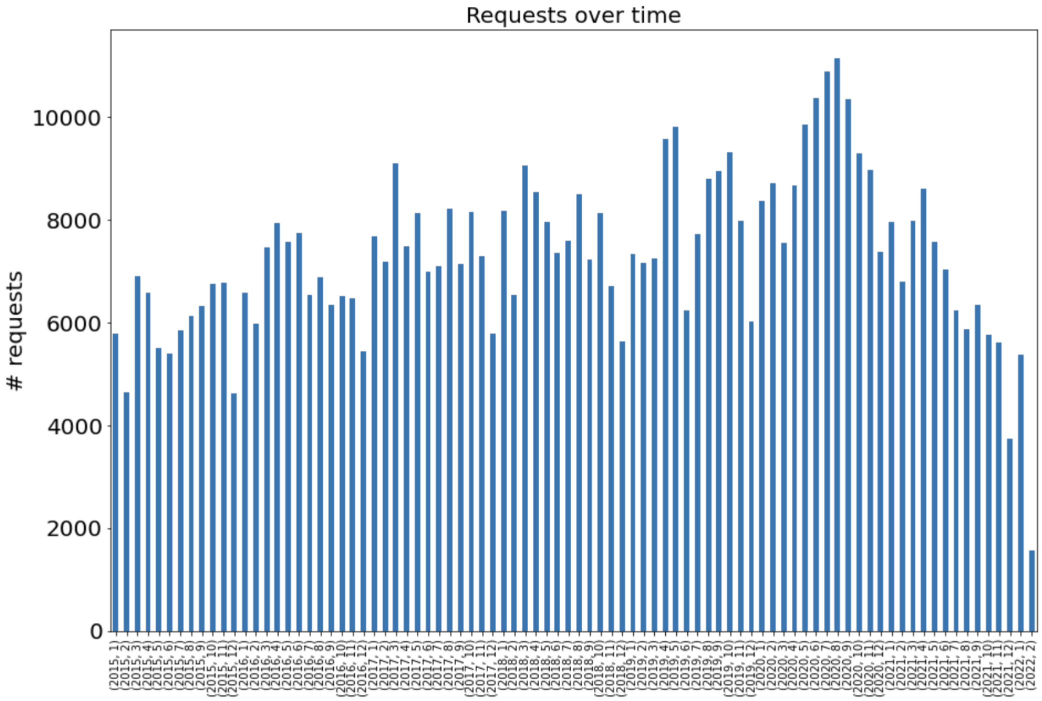


Figure 1—Number of requests (Pedidos) over time

As we are particularly interested in the requests made towards the "access to information" we use the column related to the subject of the request to see its trend over time as shown in Figure 2 and 3. What is clear is that, despite a downtrend in the total number of requests, requests to access to information are increasing both in absolute and relative terms.

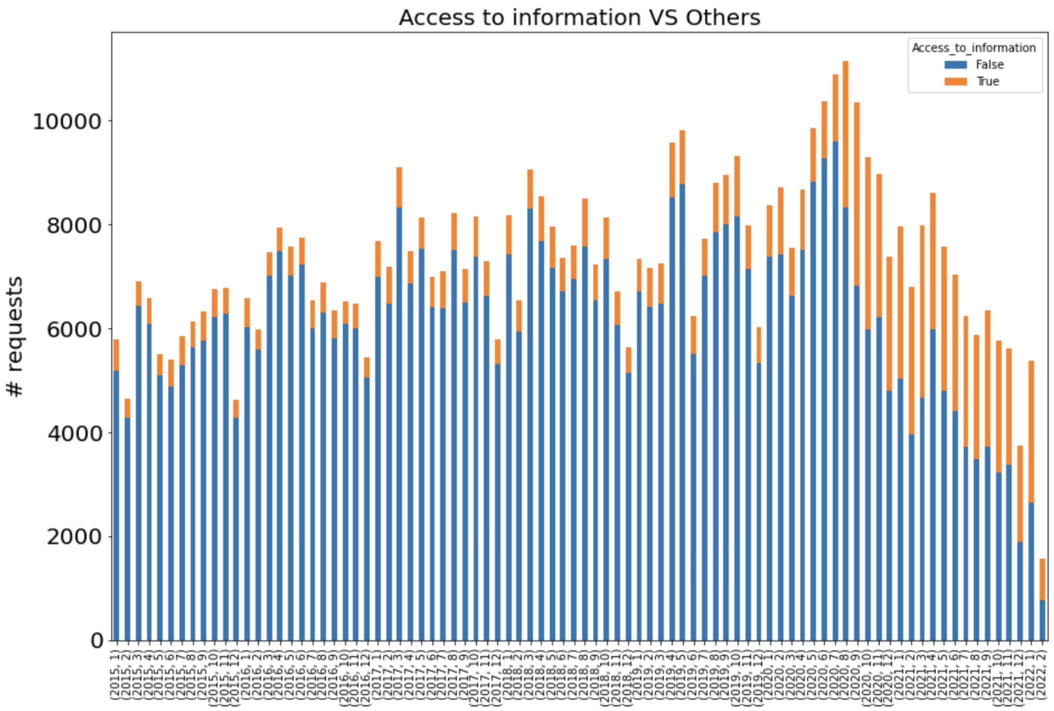


Figure 2—Absolute volume of access to information versus the rest of other subjects

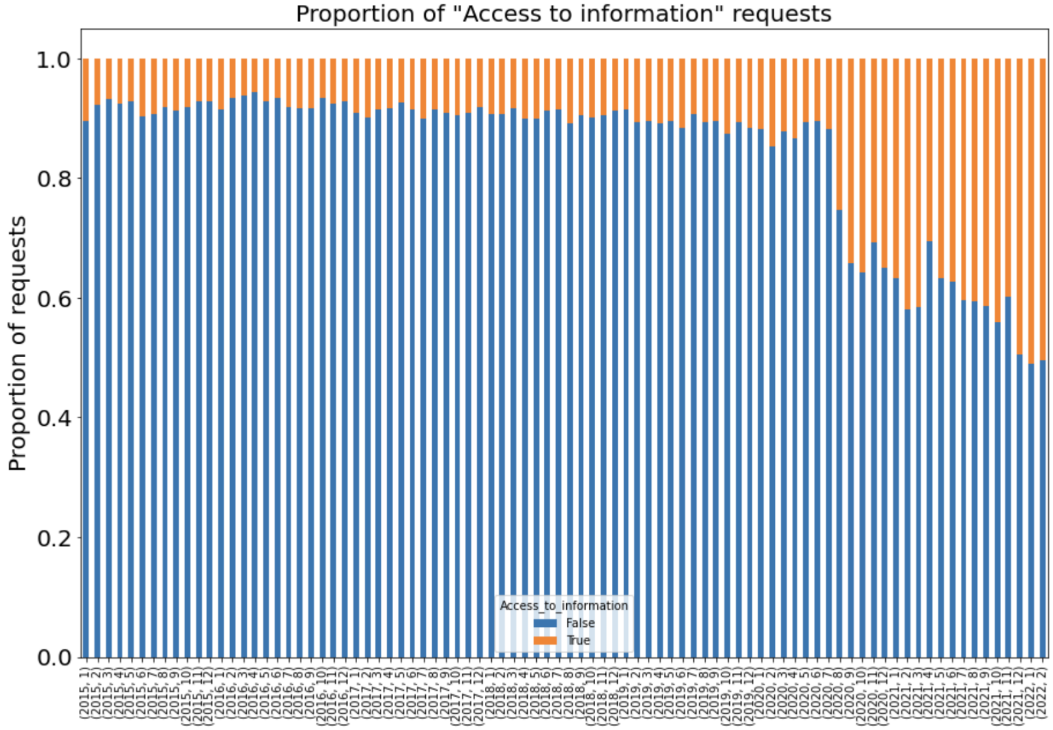


Figure 3—Proportional trend for access to information

Once we zoom in this category we can see other interesting trends as well. By analyzing the evolution of “access to information” treatment by the agency we can reveal that there has been an increase in the share of denials to this type of request as Figures 4, 5 and 6 show. The column “Decisao” is used in these cases.

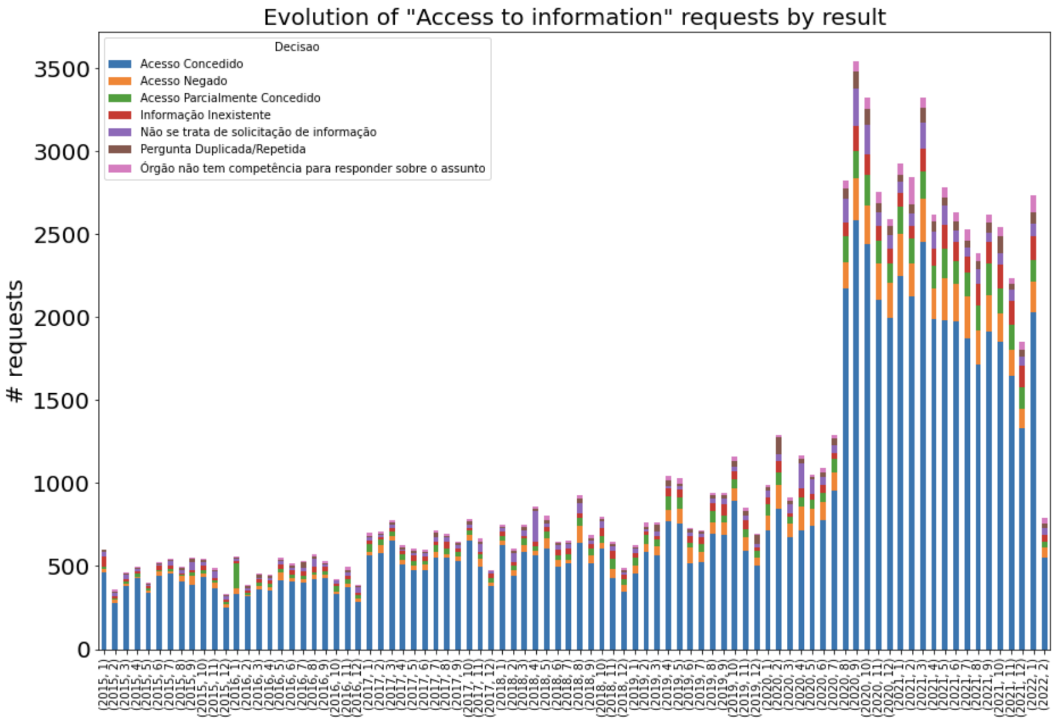


Figure 4 —Trends based on the treatment given by agency

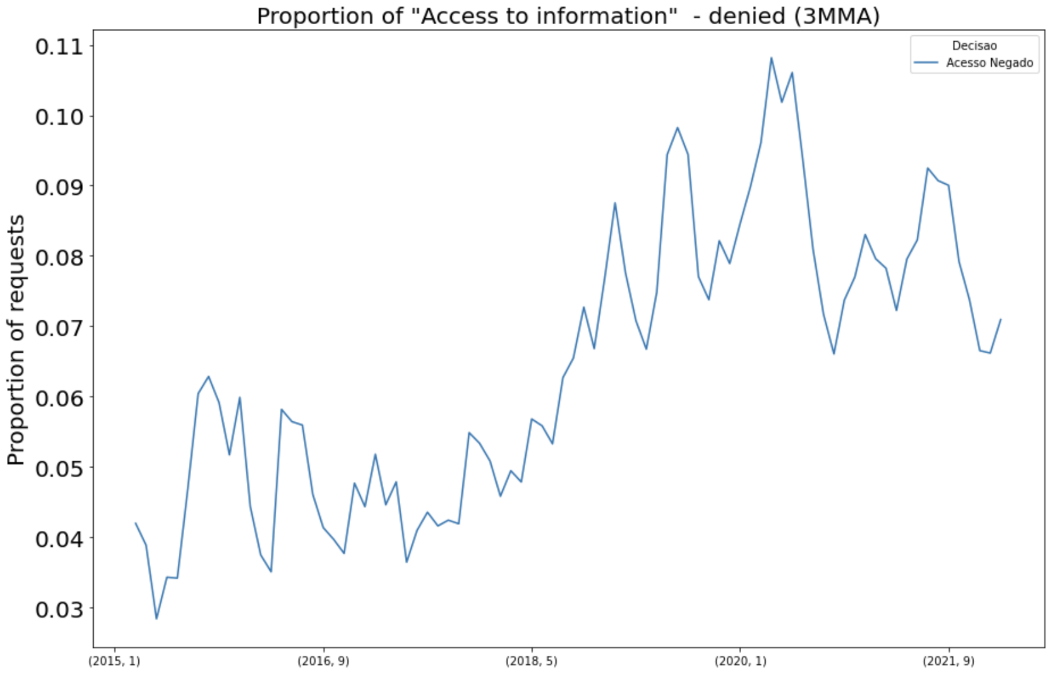


Figure 4—Rise of requests for information being denied

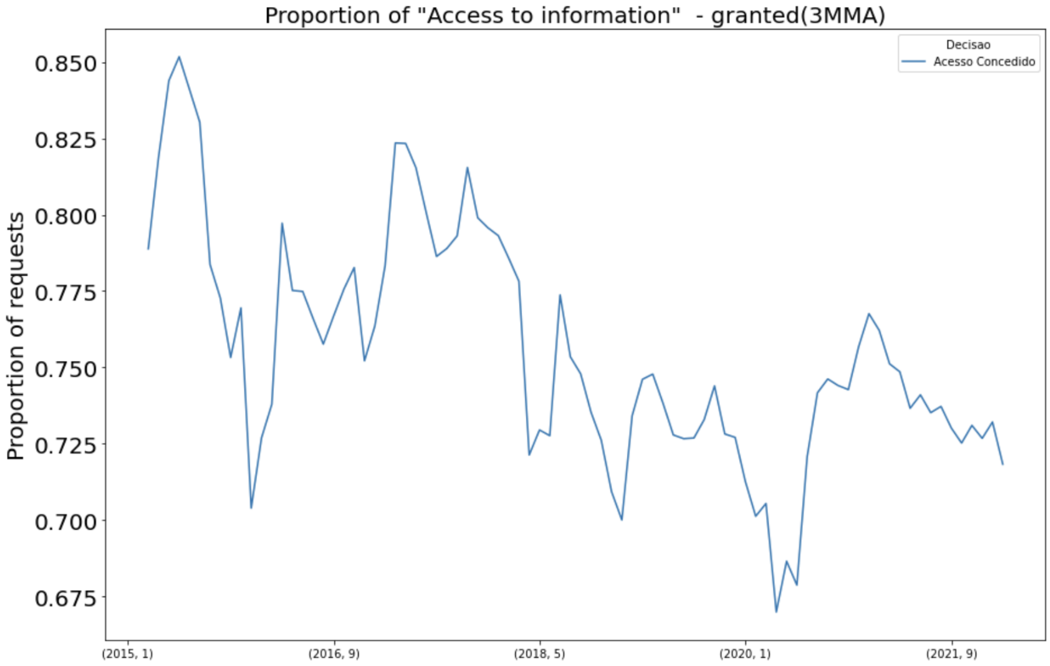


Figure 5—Decrease in "access to information" requests being granted

The agency also includes details on the reasons for denying the information, based on a set of categories. We can visualize these trends based on the column "EspecificacaoDecisao". In Figures 6 and 7 we can see such reasons for all the denials for "access to information".

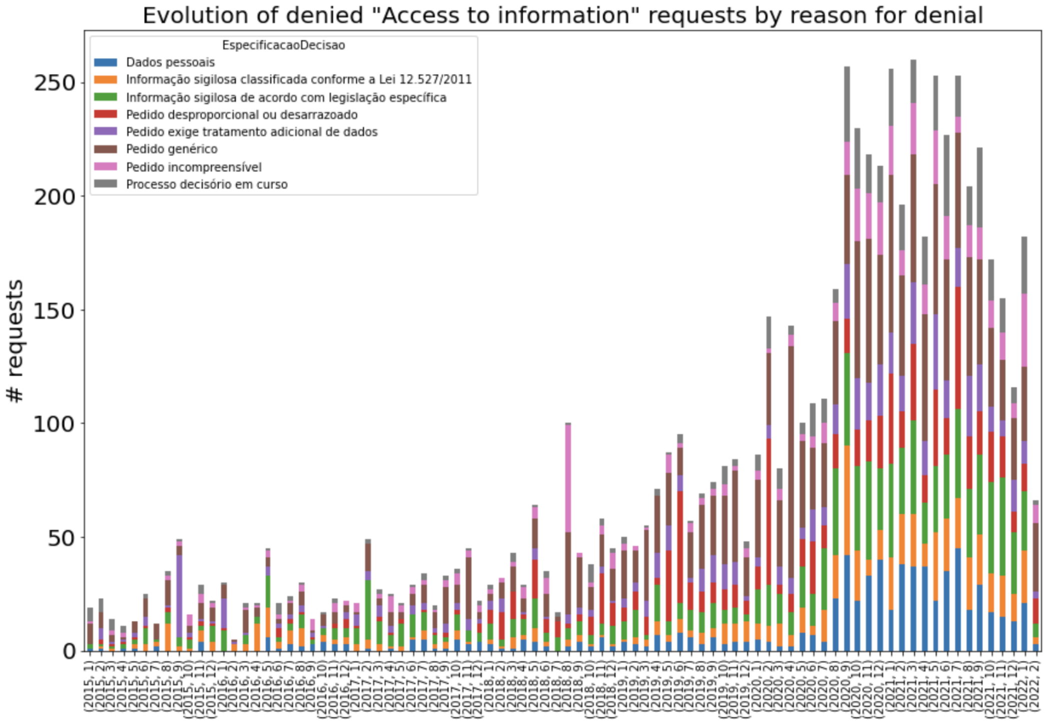


Figure 6—Trends in “Access to information” denials

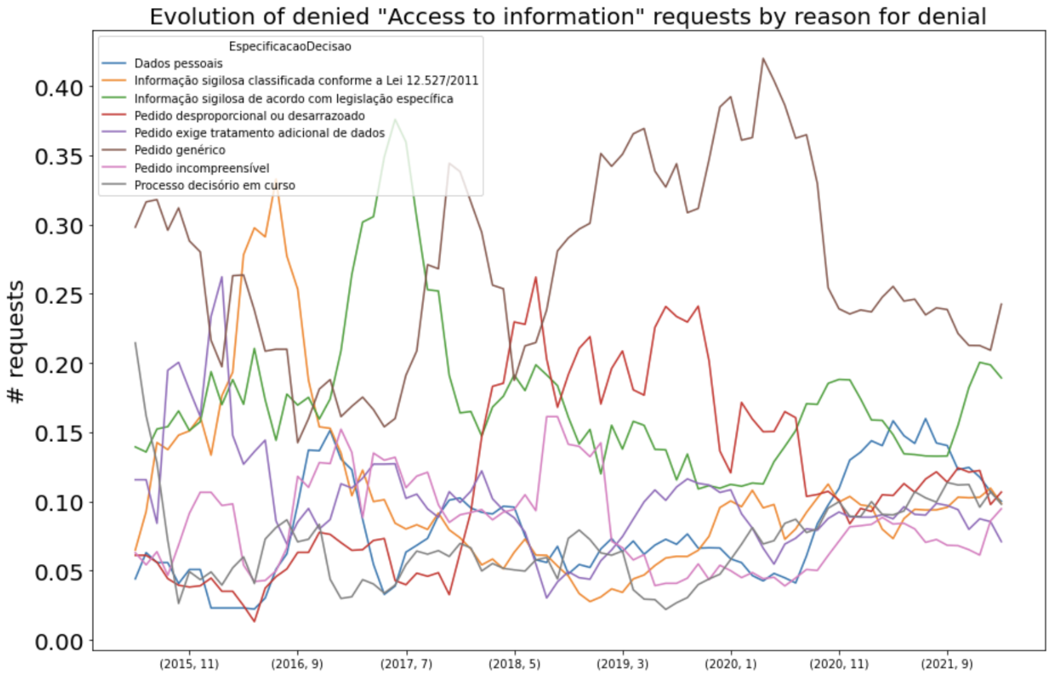


Figure 7—Share of denials by reason for "access to information" request

3.2 Reflections

Upon analyzing the texts on the requests and responses we feel that the taxonomy used for classifying them can feel arbitrary and hide some important insights.

To start addressing this problem we first create an independent taxonomy to categorize the requests. In trying to remove biases in this new classification we decided to use unsupervised machine learning models which can be later analyzed by specialists.

We also see an opportunity to create a service where anyone can test their queries and check for the probability of being denied. Such a service could be hosted on a public portal and make use of an API to make such inferences.

Details of both solutions can be found in the next sections.

4 UNCOVERING TOPICS

In this section we demonstrate a method for trying to understand what are the topics that make up the data.

4.1 Scope

It is our particular interest to understand more about the requests where claims over protection of personal data were used to deny access to information and eventually misused, after filtering only such requests we can identify which topics lie within this group.

We use the keywords ‘lgpd', '13.709' and 'lpgd' to filter the claims, resulting in 1679 claims.

4.2 Algorithms and methods

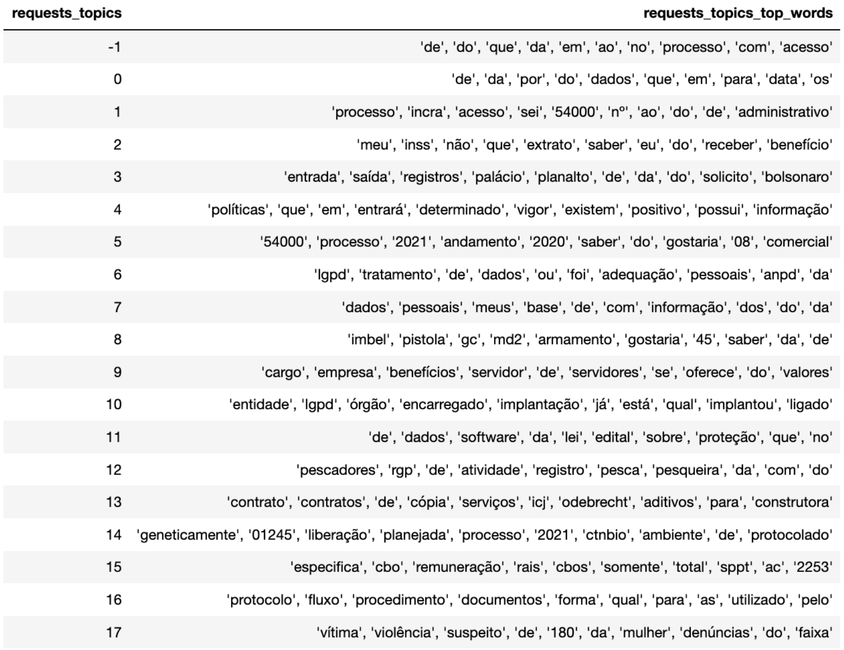
In order to create topics in an unsupervised manner we rely on a library called BERTopic (Grootendorst, n.d.), which uses sentence embeddings trained on BERT (devlin, 2018, bert) architecture to create clusters.

The method is based on a three step approach:

1) **Sentence embeddings:** The first step is to generate embeddings for the sentences in our problem. The embeddings are vectors that try to represent the semantics of the phrase in numbers. For the purposes of this project we are treating each request as one single phrase. The embeddings are generated using a pretrained model (*Pretrained Models — Sentence-Transformers Documentation*, n.d.) and supports multiple languages, including Portuguese.
2) **Dimensionality reduction:** Many clustering algorithms have a better performance if dealing with a reduced number of dimensions, for this purpose BERTopic applies UMAP (mcinnes, 2018, umap) before proceeding to the next step.
3) **Clustering:** The final step is to create the clusters based on the reduced embeddings. HDBSCAN (malzer, 2020, hybrid) is used as it does not require a predefined number of clusters to be set as a parameter.

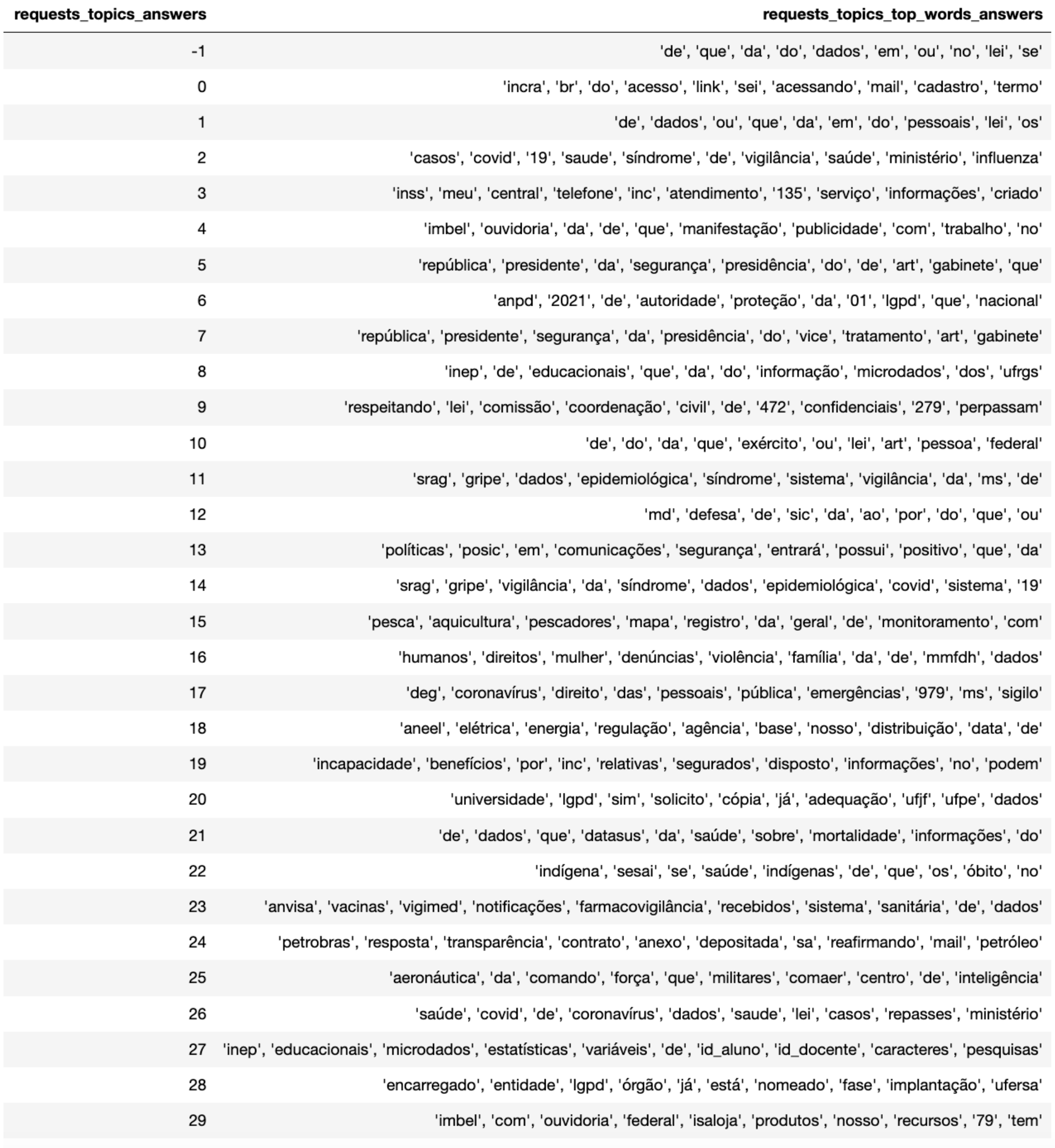
4.3 Results

The Figures 8 and 9 show the topics formed along with the words that are the most important to distinguish one group from another. Topic -1 represent unassigned requests.



requests_topics	requests_topics_top_words
-1	'de', 'do', 'que', 'da', 'em', 'ao', 'no', 'processo', 'com', 'acesso'
0	'de', 'da', 'por', 'do', 'dados', 'que', 'em', 'para', 'data', 'os'
1	'processo', 'incra', 'acesso', 'sei', '54000', 'nº', 'ao', 'do', 'de', 'administrativo'
2	'meu', 'inss', 'não', 'que', 'extrato', 'saber', 'eu', 'do', 'receber', 'benefício'
3	'entrada', 'saída', 'registros', 'palácio', 'planalto', 'de', 'da', 'do', 'solicito', 'bolsonaro'
4	'políticas', 'que', 'em', 'entrará', 'determinado', 'vigor', 'existem', 'positivo', 'possui', 'informação'
5	'54000', 'processo', '2021', 'andamento', '2020', 'saber', 'do', 'gostaria', '08', 'comercial'
6	'lgpd', 'tratamento', 'de', 'dados', 'ou', 'foi', 'adequação', 'pessoais', 'anpd', 'da'
7	'dados', 'pessoais', 'meus', 'base', 'de', 'com', 'informação', 'dos', 'do', 'da'
8	'imbel', 'pistola', 'gc', 'md2', 'armamento', 'gostaria', '45', 'saber', 'da', 'de'
9	'cargo', 'empresa', 'benefícios', 'servidor', 'de', 'servidores', 'se', 'oferece', 'do', 'valores'
10	'entidade', 'lgpd', 'órgão', 'encarregado', 'implantação', 'já', 'está', 'qual', 'implantou', 'ligado'
11	'de', 'dados', 'software', 'da', 'lei', 'edital', 'sobre', 'proteção', 'que', 'no'
12	'pescadores', 'rgp', 'de', 'atividade', 'registro', 'pesca', 'pesqueira', 'da', 'com', 'do'
13	'contrato', 'contratos', 'de', 'cópia', 'serviços', 'icj', 'odebrecht', 'aditivos', 'para', 'construtora'
14	'geneticamente', '01245', 'liberação', 'planejada', 'processo', '2021', 'ctnbio', 'ambiente', 'de', 'protocolado'
15	'especifica', 'cbo', 'remuneração', 'rais', 'cbos', 'somente', 'total', 'sppt', 'ac', '2253'
16	'protocolo', 'fluxo', 'procedimento', 'documentos', 'forma', 'qual', 'para', 'as', 'utilizado', 'pelo'
17	'vítima', 'violência', 'suspeito', 'de', '180', 'da', 'mulher', 'denúncias', 'do', 'faixa'

Figure 8 —Topics and their most distinguishing words (request text)



requests_topics_answers	requests_topics_top_words_answers
-1	'de', 'que', 'da', 'do', 'dados', 'em', 'ou', 'no', 'lei', 'se'
0	'incra', 'br', 'do', 'acesso', 'link', 'sei', 'acessando', 'mail', 'cadastro', 'termo'
1	'de', 'dados', 'ou', 'que', 'da', 'em', 'do', 'pessoais', 'lei', 'os'
2	'casos', 'covid', '19', 'saude', 'síndrome', 'de', 'vigilância', 'saúde', 'ministério', 'influenza'
3	'inss', 'meu', 'central', 'telefone', 'inc', 'atendimento', '135', 'serviço', 'informações', 'criado'
4	'imbel', 'ouvidoria', 'da', 'de', 'que', 'manifestação', 'publicidade', 'com', 'trabalho', 'no'
5	'república', 'presidente', 'da', 'segurança', 'presidência', 'do', 'de', 'art', 'gabinete', 'que'
6	'anpd', '2021', 'de', 'autoridade', 'proteção', 'da', '01', 'lgpd', 'que', 'nacional'
7	'república', 'presidente', 'segurança', 'da', 'presidência', 'do', 'vice', 'tratamento', 'art', 'gabinete'
8	'inep', 'de', 'educacionais', 'que', 'da', 'do', 'informação', 'microdados', 'dos', 'ufrgs'
9	'respeitando', 'lei', 'comissão', 'coordenação', 'civil', 'de', '472', 'confidenciais', '279', 'perpassam'
10	'de', 'do', 'da', 'que', 'exército', 'ou', 'lei', 'art', 'pessoa', 'federal'
11	'srag', 'gripe', 'dados', 'epidemiológica', 'síndrome', 'sistema', 'vigilância', 'da', 'ms', 'de'
12	'md', 'defesa', 'de', 'sic', 'da', 'ao', 'por', 'do', 'que', 'ou'
13	'políticas', 'posic', 'em', 'comunicações', 'segurança', 'entrará', 'possui', 'positivo', 'que', 'da'
14	'srag', 'gripe', 'vigilância', 'da', 'síndrome', 'dados', 'epidemiológica', 'covid', 'sistema', '19'
15	'pesca', 'aquicultura', 'pescadores', 'mapa', 'registro', 'da', 'geral', 'de', 'monitoramento', 'com'
16	'humanos', 'direitos', 'mulher', 'denúncias', 'violência', 'família', 'da', 'de', 'mmfdh', 'dados'
17	'deg', 'coronavírus', 'direito', 'das', 'pessoais', 'pública', 'emergências', '979', 'ms', 'sigilo'
18	'aneel', 'elétrica', 'energia', 'regulação', 'agência', 'base', 'nosso', 'distribuição', 'data', 'de'
19	'incapacidade', 'benefícios', 'por', 'inc', 'relativas', 'segurados', 'disposto', 'informações', 'no', 'podem'
20	'universidade', 'lgpd', 'sim', 'solicito', 'cópia', 'já', 'adequação', 'ufjf', 'ufpe', 'dados'
21	'de', 'dados', 'que', 'datasus', 'da', 'saúde', 'sobre', 'mortalidade', 'informações', 'do'
22	'indígena', 'sesai', 'se', 'saúde', 'indígenas', 'de', 'que', 'os', 'óbito', 'no'
23	'anvisa', 'vacinas', 'vigimed', 'notificações', 'farmacovigilância', 'recebidos', 'sistema', 'sanitária', 'de', 'dados'
24	'petrobras', 'resposta', 'transparência', 'contrato', 'anexo', 'depositada', 'sa', 'reafirmando', 'mail', 'petróleo'
25	'aeronáutica', 'da', 'comando', 'força', 'que', 'militares', 'comaer', 'centro', 'de', 'inteligência'
26	'saúde', 'covid', 'de', 'coronavírus', 'dados', 'saude', 'lei', 'casos', 'repasses', 'ministério'
27	'inep', 'educacionais', 'microdados', 'estatísticas', 'variáveis', 'de', 'id_aluno', 'id_docente', 'caracteres', 'pesquisas'
28	'encarregado', 'entidade', 'lgpd', 'órgão', 'já', 'está', 'nomeado', 'fase', 'implantação', 'ufersa'
29	'imbel', 'com', 'ouvidoria', 'federal', 'isaloja', 'produtos', 'nosso', 'recursos', '79', 'tem'

Figure 8—Topics and their most distinguishing words (response text)

Not all topics created by the method make sense, but help in getting an idea about what is being discussed and can trigger further investigations by specialists.

5 PROBABILITY OF DENIAL SERVICE

As a way to help citizens to identify if their requests have a high chance of being rejected by the agency we propose the deployment of a system that will receive a free text and provide the probability of being rejected

4.1 Scope

To deploy such a system, a Machine learning algorithm is trained and deployed in the cloud. An API is made available for connecting to a portal and making it accessible to the general public.

4.2 Algorithms and methods

The prediction is done using the 'BERT Base Multilingual Cased' model (pretrained) which is fine tuned to predict the two classes we are interested in denied (0) or granted (1).

The training and testing follows standard procedures as and is evaluated using accuracy and recall.

4.2.1 Data preprocessing

The detailed preprocessing steps can be found on the jupyter notebook named 'setup' under notebooks folder. In this section we give a high level overview of the choices made.

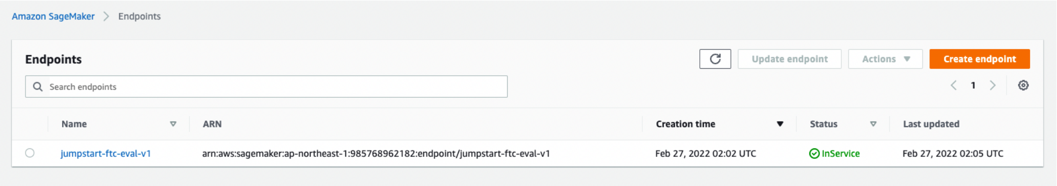
a) Download data from official portal
b) Aggregate yearly files into one single 'Pedidos' file
c) Cast the date fields into date format and remove all rows that came out as 'None' (this will remove malformed data which are very few in total)
d) Create a target label based on 'Decisao' field. So we remove the 'partially granted', 'information non existent', 'duplicated request' and 'The agency does not have the capability to answer request values'. Furthermore combine 'Denied' with 'It is not a request for information.
e) The remaining labels should be two classes which should be encoded with 0 (denied) or 1 (granted)
f) Remove all columns except for 'DetalhamentoSolicitacao' and the target column created in the step before.

g) Subset the data into a training and a test set. Push the training set without index or headers to an s3 bucket where it would be training. The files has to be named data.csv

Note that it is not necessary to preprocess the verbatim itself as BERT models expect the text in its raw form.

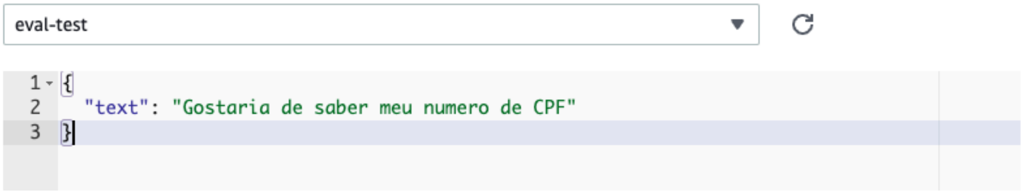
4.2.2 Deployment

The deployment is done using AWS, where an API is made available.



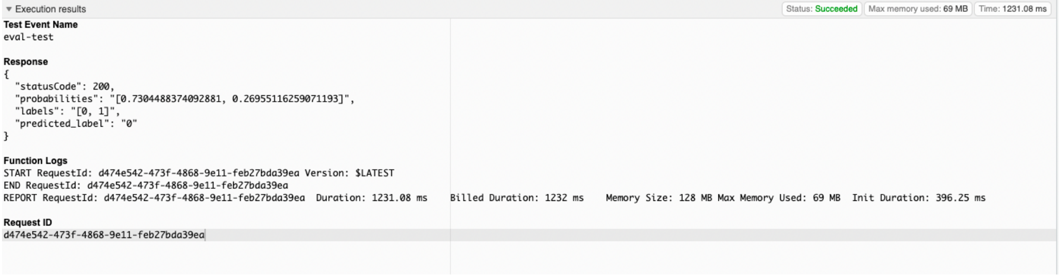
	Name	ARN	Creation time	Status	Last updated
○	jumpstart-ftc-eval-v1	arn:aws:sagemaker:ap-northeast-1:985768962182:endpoint/jumpstart-ftc-eval-v1	Feb 27, 2022 02:02 UTC	✓ InService	Feb 27, 2022 02:05 UTC

Figure 9—Deployed endpoint



```
{
  "text": "Gostaria de saber meu numero de CPF"
}
```

Figure 10 —Example on how to send a request



```
Test Event Name
eval-test

Response
{
  "statusCode": 200,
  "probabilities": "[0.7304488374092881, 0.26955116259071193]",
  "labels": "[0, 1]",
  "predicted_label": "0"
}

Function Logs
START RequestId: d474e542-473f-4868-9e11-feb27bda39ea Version: $LATEST
END RequestId: d474e542-473f-4868-9e11-feb27bda39ea
REPORT RequestId: d474e542-473f-4868-9e11-feb27bda39ea  Duration: 1231.08 ms    Billed Duration: 1232 ms    Memory Size: 128 MB Max Memory Used: 69 MB  Init Duration: 396.25 ms

Request ID
d474e542-473f-4868-9e11-feb27bda39ea
```

Figure 11—Sample response

6 BENCHMARK

The novelty of the services proposed in this project makes it difficult to establish a benchmark with which it is possible to make a comparison. Nevertheless we want to commit to a standard of quality to assess if the API is working reasonably well, so we aimed for an 0.7+ F1 score.

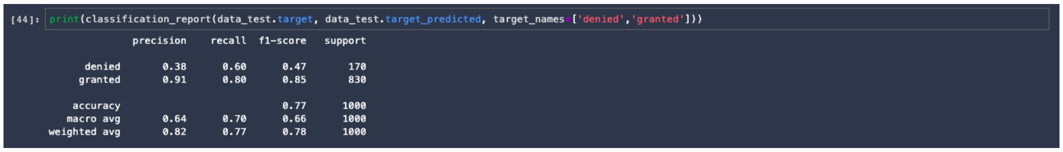
7 METRICS

As mentioned in the previous section the metric to be used in this project in F1 score, although we use other metrics to support our conclusions.

8 IMPROVEMENTS

Figure 12 shows the performance on a test set. It is clear that it is well below the standard that it was hoped for in the beginning, particularly when it comes to the 'denied' class.

It was trained only on 10000 samples (balanced) and standard hyperparatemers



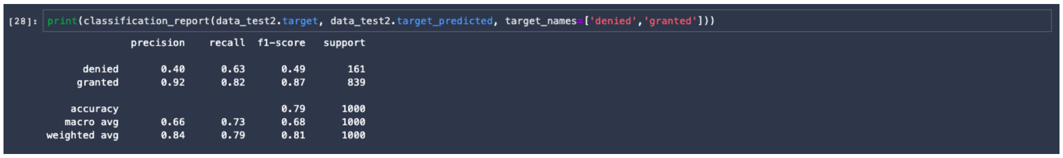
```
[44]: print(classification_report(data_test.target, data_test.target_predicted, target_names=['denied','granted']))
                  precision    recall  f1-score   support

          denied       0.38      0.60      0.47       170
         granted       0.91      0.80      0.85       830

        accuracy                           0.77      1000
       macro avg       0.64      0.70      0.66      1000
    weighted avg       0.82      0.77      0.78      1000
```

Figure 12—Performance of model on test set

Figure 13 shows an improvement in all metrics. The changes applied were mainly on doubling the data training set to 20000 samples (balanced) and fine tuning the hyperparameters, utilizing 3 epochs, leaning rate of 0.00002 and batch size of 25.



```
[28]: print(classification_report(data_test2.target, data_test2.target_predicted, target_names=['denied','granted']))
                  precision    recall  f1-score   support

          denied       0.40      0.63      0.49       161
         granted       0.92      0.82      0.87       839

        accuracy                           0.79      1000
       macro avg       0.66      0.73      0.68      1000
    weighted avg       0.84      0.79      0.81      1000
```

Figure 12—Performance of model on test set after improvements

9 REFERENCES

9.1 Papers

(devlin, 2018, bert) Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018) (chen, 2016, xgboost) Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016

(malzer, 2020, hybrid) Malzer, Claudia, and Marcus Baum. "A hybrid approach to hierarchical density-based cluster selection." *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 2020

(mcinnes, 2018, umap) McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." *arXiv preprint arXiv:1802.03426* (2018)

9.2 Websites

Acesso a information. (n.d.). Governo Federal. Retrieved February 22, 2022, from

https://www.gov.br/acessoainformacao/pt-br

Grootendorst, M. (n.d.). *BERTopic*. GitHub. Retrieved February 22, 2022, from

https://github.com/MaartenGr/BERTopic

Pretrained Models — Sentence-Transformers documentation. (n.d.).

Sentence-Transformers. Retrieved February 22, 2022, from

https://www.sbert.net/docs/pretrained_models.html