Trabalho Final - Análise dos Discursos da CPI da Pandemia

Estatística e Ciência de Dados - PUC-Rio Verão 2022

Bernardo Duque

27/02/2022

Contents

1 Objetivo **1**

2 Dados Utilizados **2**

3 Metodologia **2**

3.1 Tratamento das strings 2

3.2 Machine Learning 3

3.3 Métricas de Validação 4

3.4 Criação de Base de Dados 5

4 Estatísticas Descritivas **5**

4.1 Nuvens de palavras 5

4.2 Dados dos Senadores 6

4.3 Dados dos Discursos 6

4.4 Dados Covid 11

5 Conclusão e Limitações **13**

1 Objetivo

Este trabalho teve como objetivo analisar os discursos e a composição da Comissão Parlamentar de Inquérito (CPI) da Pandemia, que teve início em 27 de abril de 2021 e fim em 26 de outubro do mesmo ano. A comissão teve ampla divulgação midiática e alto acompanhamento por parte da população[1]. Sua relevância se deu principalmente pelo descaso do Governo Federal no combate à crise sanitária instalada no país e no mundo.

Para isso, foi criado um modelo de *Machine Learning (ML)* para automatizar a análise, indicando se o discurso era favorável ou contra atuação da União durante a pandemia. Também foram construídas estatísticas

[1]De acordo com dados do Instituto DataSenado, 73% da população tomou conhecimento da CPI. Fonte: https://www12.senado.leg.br/noticias/materias/2021/07/19/datasenado-73-dos-brasileiros-conhecem-a-cpi-da-pandemia

descritivas e mapas com a composição dos participantes da CPI, com o output do modelo de ML e com o cruzamento dos dados de covid no país.

Vale destacar que este trabalho não tem a pretensão de ser conclusivo e apresenta diversas limitações que serão tratadas na última seção. O modelo não performou como esperado, portanto a ideia aqui é mostrar que a análise é possível, mas melhorias são necessárias. Ou seja, servir como uma *prova de conceito* de que podemos utilizar os dados do discursos tratados por um modelo de ML e relacionar o andamento da pandemia em cada estado com os discursos contrários e favoráveis por parte dos senadores de origem nestes locais.

2 Dados Utilizados

Para realizar a análise, foram utilizadas inicialmente 3 bases de dados, a saber:

- **Base com discursos da CPI da Pandemia**
  - Principal base utilizada neste trabalho. Contém todas as falas no decorrer da CPI, tanto de senadores quanto de depoentes, incluindo tambem falas de ordem e de reuniões de requerimentos
- **Dados de covid**
  - Base com dados acumulados da pandemia, com acumulação diária. Inclui número de casos, número de mortes, taxas por cem mil habitantes e a unidade de análise está no nível dos estados
- **Dados dos candidatos**
  - Base com informações dos candidatos, como idade, instrução, profissão, gênero, entre outras

As bases podem ser acessadas diretamente pelos link acima ou clicando aqui.

3 Metodologia

3.1 Tratamento das strings

Como visto na seção anterior, a principal base utilizada foi a relativa aos discursos da CPI da Pandemia. Por se tratar de uma base cujos dados principais estavam no formato de texto (string), foi necessário um tratamento considerável dos dados para convertê-los em informação. Foram analisadas apenas as falas dos senadores; as dos depoentes foram filtradas. Vale ressaltar que foram considerados todos os senadores que fizeram algum discurso durante as sessões, mesmo caso não fosse integrante da comissão.

De forma geral, o processo adotado para tratar as strings foi a chamada tokenização. Pegou-se todas as palavras contidas em cada discurso e cada uma foi transformada em uma linha com as demais informações duplicadas do discurso. Em seguida, foram retiradas as chamadas *stop words*, que são palavras que não acrescentam informação, mas geram ruído nas análises e aumentam o custo de processamento do algoritmo.

Desses primeiros passos, já foi possível extrair alguma informação, por meio de nuvens de palavras, que indicam as palavras que tiveram maior frequência durante os discursos. Uma foi gerada após esses passos já explicados, enquanto a segunda foi feita após o processo de *stemming*. Isto é, foram agrupadas palavras com bases em seus radicais; isso permite reduzir o número de palavras analisadas, focando em seus significados, ou seja, reduzir a quantidade de variáveis sem perder informação.

Em seguida, foram retiradas as palavras com baixa frequência, que apareciam menos de 21 vezes e correspodiam a 79% de todas as palavras. Por fim, cada palavra "stemizada" foi transformada em uma coluna (variável), permitindo a aplicação do modelo de ML.

3.2 Machine Learning

Anterior a isso, porém, foi gerada uma amostra aleatória replicável de 1000 observações da base original com os discursos[2]. As falas foram classificadas como:

- **Contra a atuação do Governo Federal na pandemia**
  - Inclui menções criticando a atuação do governo de forma geral, por meio da crítica a seus ministérios ou à própria pessoa do Presidente da República
  - Exemplo:

    "O BRASIL NAO E PARIA, O BRASIL FOI MUITO PIOR: EM VEZ DE PARIA, O SENHOR COLOCOU O BRASIL NA POSICAO DE IRRELEVANCIA! E EU NAO ACEITO O MEU PAIS SER UM PAIS IRRELEVANTE! NAO ACEITO QUE ISSO ACONTECA! PARA CONCLUIR, SR. PRESIDENTE, O MAIOR VEXAME QUE NOS JA PASSAMOS NA VIDA..."

- **A favor da atuação do Governo Federal na pandemia**
  - Inclui menções que louvam a atuação do governo de forma geral, por meio de apologia a medicamentos do chamado Kit Covid e defesas às críticas de outros senadores
  - Exemplo:

    "SEGUINDO, NO DIA 11 DE JANEIRO, AS 16H, O MINISTERIO DA SAUDE E O ESTADO DO AMAZONAS INSTAURARAM O CENTRO INTEGRADO DE COORDENACAO E CONTROLE PARA AUXILIAR NA SOLUCAO DE TODAS AS DEMANDAS ESTADUAIS ADVINDAS DA CRISE EM MANAUS. ENCERRO, SR. PRESIDENTE, COM ESSAS LINHAS GERAIS. PORTANTO, O MINISTERIO DA SAUDE, COM O APOIO DO COMANDO MILITAR DA AMAZONIA, DE FORMA EFETIVA, AUXILIOU NA SOLUCAO DAS DEMANDAS SANITARIAS, ALEM DE ARTICULAR AS ACOES NO AMBITO DOS DEMAIS MINISTERIOS DO GOVERNO FEDERAL. MUITO OBRIGADO, E EU PECO AS MINHAS ESCUSAS AO MEU QUERIDO COMPANHEIRO SENADOR ALESSANDRO VIEIRA." "'

- **Neutras ou de questões de ordem**
  - Inclui falas que não criticam nem defendem o Governo Federal, falas sobre requerimentos, perguntas sem juízo de valor ou questões de ordem corriqueiras de uma CPI
  - Exemplo:

    "POR FAVOR, SENADOR."

Com a classificação manual da amostra e com o tratamento das strings, finalmente foi possível aplicar um modelo de ML. A partir da classificação manual, foram separadas uma base de treino e outra de teste para o modelo, na proporção 65/35 respectivamente. Ou seja, 650 falas foram utilizadas como treino do modelo e 350 como teste. Importante ressaltar que elas foram alocadas aleatoriamente para cada um desses grupos.

Foram testados 2 tipos de modelos de classificação:

- **Random Forests**
- **Extreme Gradient Boosting Tree (XGBTree)**

Para ambos foi utilizada uma técnica chamada de cross-validation, que serve para evitar um *overfitting* do modelo. Intuitivamente, consiste em particionar sua base de treino e rotacionar iterativamente as partições como treino e teste, e testar a performance do modelo nessas subamostras.

[2]Foi utilizada a função set.seed()

3.3 Métricas de Validação

Após rodar os dois modelos, foram geradas as matrizes de confusão de cada um, e com base na métrica F1, foi o escolhido o modelo XGBTree, que estava melhor especificado. Infelizmente, o modelo não performou tão bem quanto o esperado, mas algumas tentativas podem ser feitas para melhorá-lo, como será discutido na seção Limitações.

A matriz de confusão apresenta as seguintes métricas:

- **Precision**: Indica a proporção de positivos verdadeiros classificados como positivos dentre o universo das observações que foram classificadas como positivas. Precision baixo que estamos classificando como positivas muitas observações falsas. É igual a 1 quando todas as classificações de positivos estão corretas.
  - Calculada por:
$$Precision = \frac{Positivos\ Verdadeiros}{Positivos\ Verdadeiros + Falsos\ Positivos}$$
- **Recall**: Mede a proporção de positivos verdadeiros classificados como positivos em relação ao universo de todos os positivos verdadeiros. Recall é igual a 1 quando todos os positivos verdadeiros foram classificados como positivos. Valor baixo indica que não estamos "peneirando" positivos verdadeiros suficientes.
  - Calculada por:
$$Recall = \frac{Positivos\ Verdadeiros}{Positivos\ Verdadeiros + Falsos\ Negativos}$$
- **F1-Score**: Adequada quando busca-se um equilíbrio entre *Precision* e *Recall* - indiferença entre falsos positivos e falsos negativos - e quando um há uma distribuição heterogênea tendendo para um grande número de negativos verdadeiros. Na realidade é a média harmônica das duas medidas.
  - Calculada por :
$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$
- **Prevalence**: Proporção de vezes que determinada classe apareceu na classificação manual

A tabela 1 apresenta a matriz de confusão do modelo utilizado:

Table 1: Matriz de Confusão do Modelo

	Precision	**Recall**	**F1**	**Prevalence**
Class: A favor	0.333	0.125	0.182	0.049
Class: Contra	0.615	0.457	0.525	0.106
Class: Neutro	0.916	0.978	0.946	0.845

Como dito anteriormente, as métricas deixam a desejar para as classificações "a favor do governo" e "contra o governo". Para a primeira, tanto o *precision* quanto o *recall* estão bem baixos, o que implica que o modelo está pegando poucos casos que deveria pegar, e os que está pegando está pegando errado. Na segunda classe, o *recall* também está bem baixo, implicando que o modelo está pegando menos casos do que deveria, mas o *precision* está num patamar mais aceitável, dando mais confiança para os discursos que estão classificados como tal.

Assim, as estatísticas encontradas devem ser analisadas com muita cautela e não devem ser tomadas como verdade, mas como um exercício para mostrar que a análise aqui empreendida é viável, desde que possua melhorias.

3.4 Criação de Base de Dados

Além do modelo de ML, também foram feitos merges com as bases dos candidatos, que possuem informações de cada um dos senadores, bem como com a base de dados acumulados de covid para a construção dos dados. Assim, este trabalho tem como produto 3 bases de dados:

- Base de dados com output do ML
  - No nível do discurso
- Base de dados com informações dos senadores participantes da pandemia
  - Inclusive número de discursos neutro, contra e a favor do governo
  - No nível do senador
  - Faltando informação para alguns senadores
- Base de dados com dados de covid e discurso
  - No nível do estado

4 Estatísticas Descritivas

4.1 Nuvens de palavras

A Figura 1 abaixo apresenta a primeira wordcloud, que foi construída após a tokenização dos discursos, enquanto a Figura 2 apresenta a wordcloud construída após o processo de *stemming*.

Fica evidente das duas figuras que as palavras mais ditas durante a comissão foram “Presidente”, “Senador” e “Senhor”. Também aparecem algumas abreviações como “Sr”(senhor) e “V”(Vossa Excelência).

Após o stemming, as palavras “vacinas”,“vacina” e possivelmente outras variações foram agrupadas sob “vacin” e ganharam maior destaque na nuvem de palavras. Entretanto, a palavra “senhor” e “senhora” nao foram agrupados sob o mesmo radical, como se esperaria. Esse problema se dá devido à função utilizada de *stemming*, que infelizmente têm menos opções em português do que em inglês. O mesmo problema ocorre para as *stopwords*, com a presença de palavras como “até” e “lá” na nuvem de palavras.

Outras palavras que merecem destaque são “ministério”, “saude”, “ministro” e “governo”, que como era de se esperar numa CPI de saúde pública aparecem com bastante frequência.

4.2 Dados dos Senadores

A Tabela 2 abaixo mostra dados descritivos dos senadores que participaram da comissão (inclusive os não-membros). Como será visto na última seção, 17 senadores de um total de 44 não tiveram seus dados encontrados.

Porém, mesmo sem essas informações, conseguimos extrair um perfil que provavelmente não está muito viesado. Os senadores participantes foram em sua maioria, sem supresa alguma, homens, brancos de meia idade. 26% dos participantes eram mulheres, 78% eram brancos e a média de idade foi de cerca de 54 anos.

4.3 Dados dos Discursos

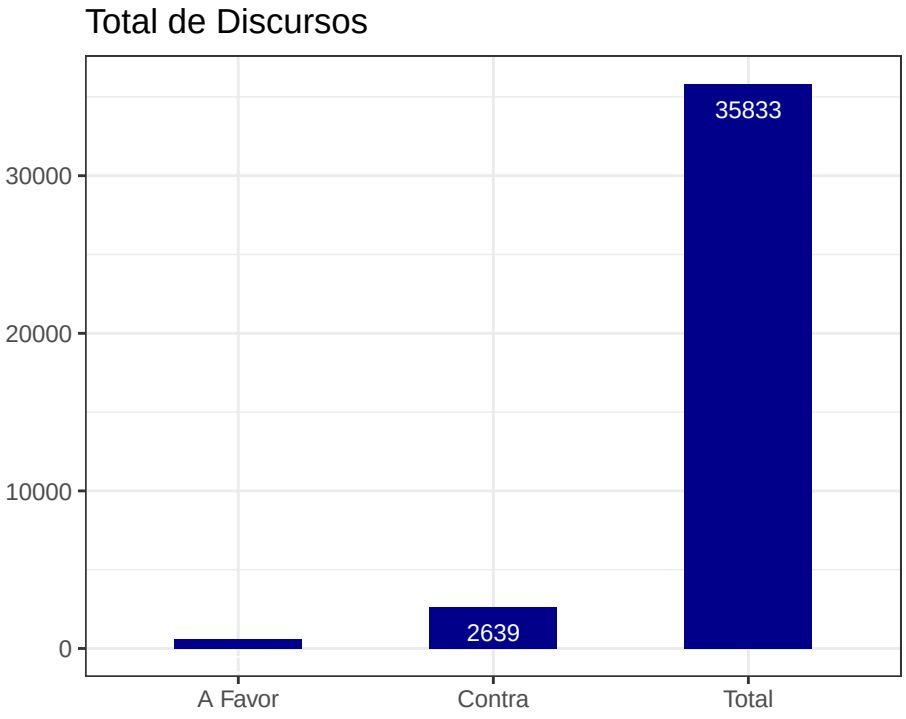
Vale ressaltar novamente que a análise dessa subseção não tem robustez dada a má especificação do modelo de ML. Porém, enquanto alguns resultados parecem duvidosos - especialmente na desagragação por senadores - a análise geral parece condizer com a impresssão anedótica.

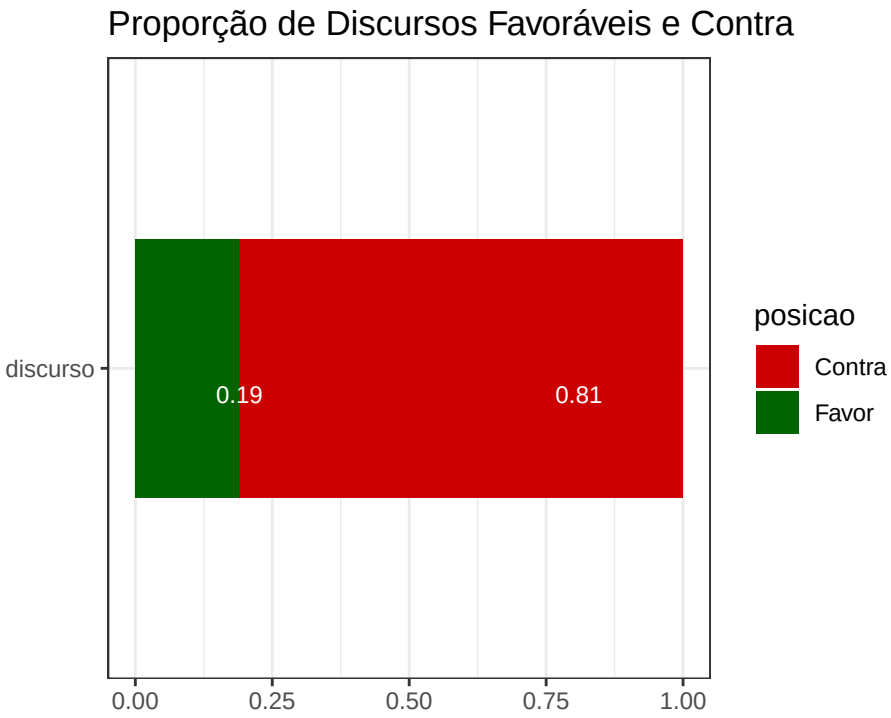
Table 2: Perfil dos Senadores

Cor ou Raça		Gênero		Idade
Cor ou Raça	**N senadores**	**Gênero**	**N de senadores**	**Média**
N/A	17	Masculino	35	53.9
branca	21	Feminino	9	
parda	4	Total	44	
preta	2			
Total	44			

4.3.1 Discursos Totais

A Figura 3 apresenta o número total de discursos por senadores durante a CPI e a Figura 4 apresenta a proporção de discursos positivos e negativos para o governo. Importante destacar que essa proporção foi construída desconsiderando os discursos neutros. A razão para isso fica evidente na discrepância exposta pela Figura 3, o que é somado ao fato de que o que importa aqui para a análise é justamente a diferença entre posições favoráveis e contrárias.

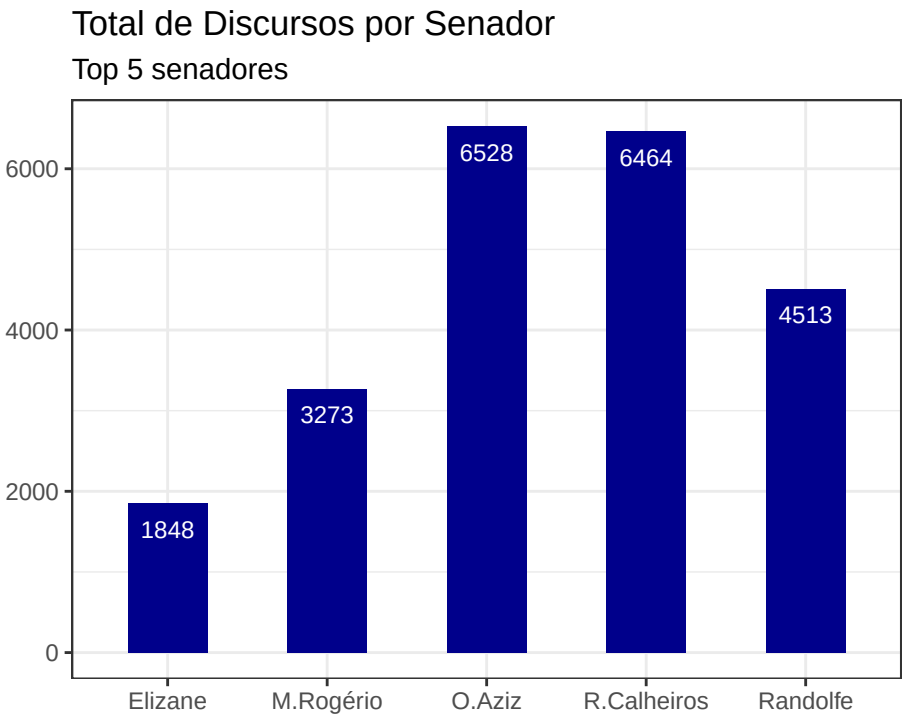




Dentre os discursos não-neutros, cerca de 81% foram críticos ao governo federal. É um número bastante alto, porém não surpreedente, dado tanto a quantidade de mortos que nosso país teve quanto a percepção da população quanto à (ausência de) resposta do governo Bolsonaro. Algo interessante de se analisar seria como a pandemia alterou o apoio dos senadores ao governo, que é de certa forma a motivação para este projeto, porém de forma mais limitada.

4.3.2 Discursos por Senadores

Conseguimos também desagregar os dados por senadores. A figura abaixo mostra os 5 senadores que mais tiveram falas durante os trabalhos da comissão.



Como era de se esperar, os 3 senadores que mais participaram foram o presiedente, o relator e o vice-presidente da comissão, respectivamente, que são os membros que têm papéis extras no Inquérito, com funções burocráticas. O senador Randolfe, além de vice-presidente, foi também o parlamentar que protocolou a criação da CPI.

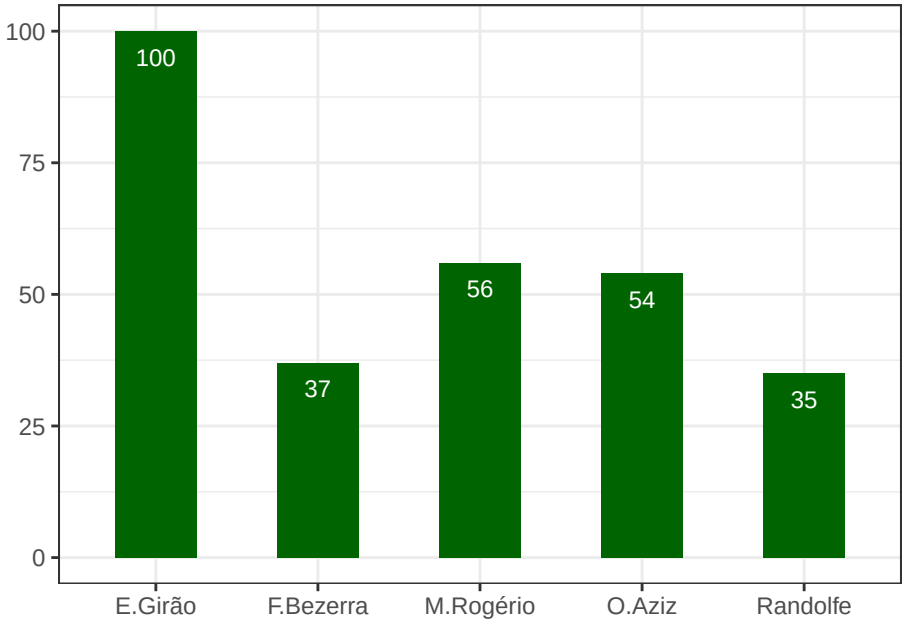
Em 4o lugar temos o senador Marcos Rogério, que ficou conhecido como ferrenho defensor do Governo Federal ao longo do período que a investigação estava ocorrendo. As duas últimas figuras dessa subseção apresentam o top 5 senadores com mais discursos favoráveis e críticos ao governo, respectivamente.

Não há surpresas em relação às posições de Eduardo Girão, Fernado Bezerra e Marcos Rogério, personagens que defenderam o presidente da república com afinco durante as sessões (apesar de Girão se dizer independente). Do outro lado, também não há surpresas, são personagens que abertamente criticaram a inação federal.

O que surpreende, porém, é a colocação de Randolfe Rodrigues e Omar Aziz, que aparecem nos dois top 5. Claro que as magnitudes são bem inferiores no caso dos discursos favoráveis ao governo, mas isso é mais um indicativo da falha do modelo de ML aqui utilizado.

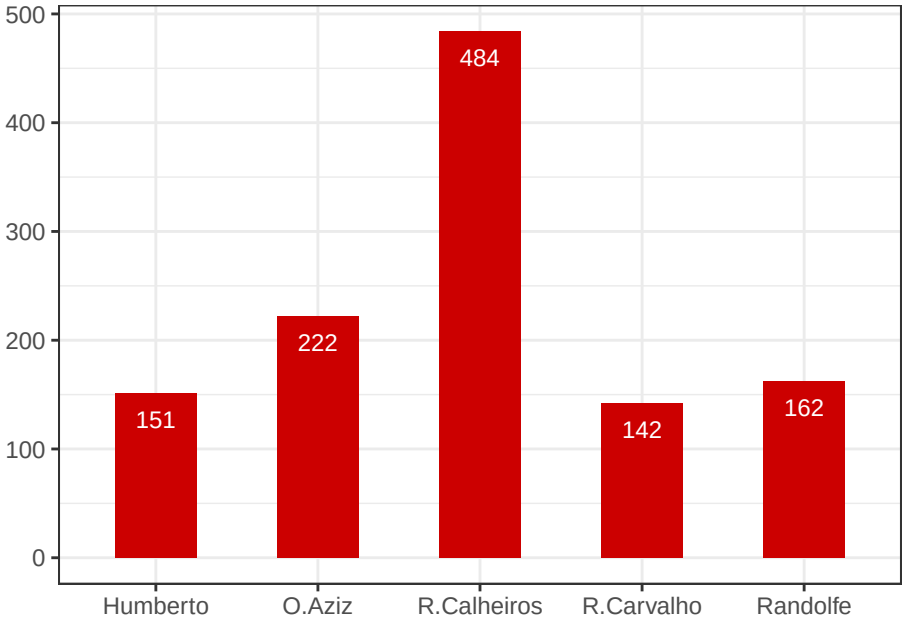
Total de Discursos a Favor do Governo por Senador

Top 5 senadores



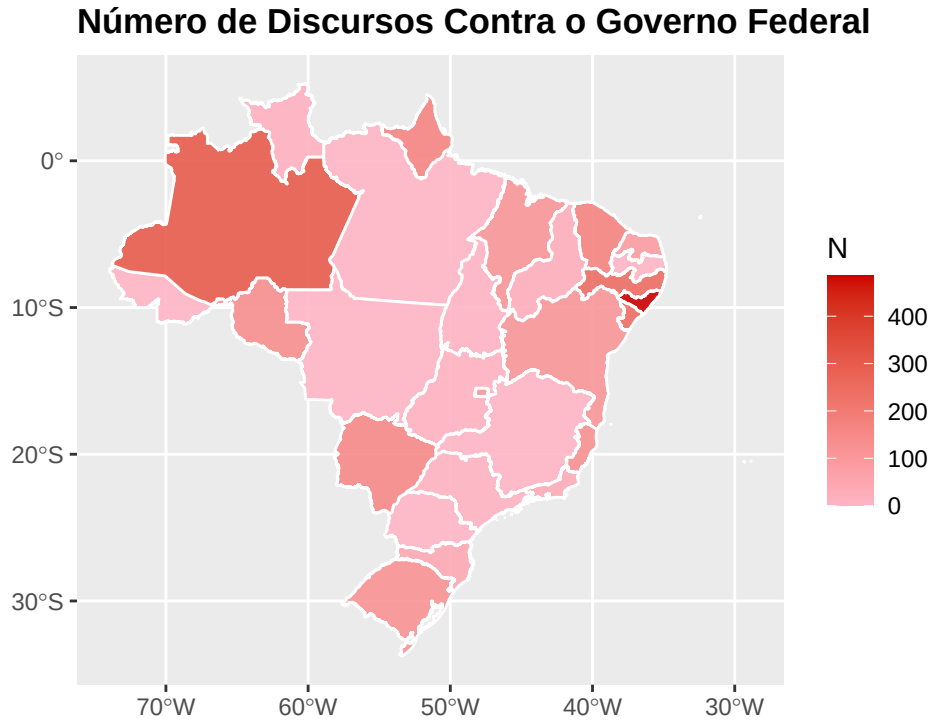
Total de Discursos Contra o Governo por Senador

Top 5 senadores

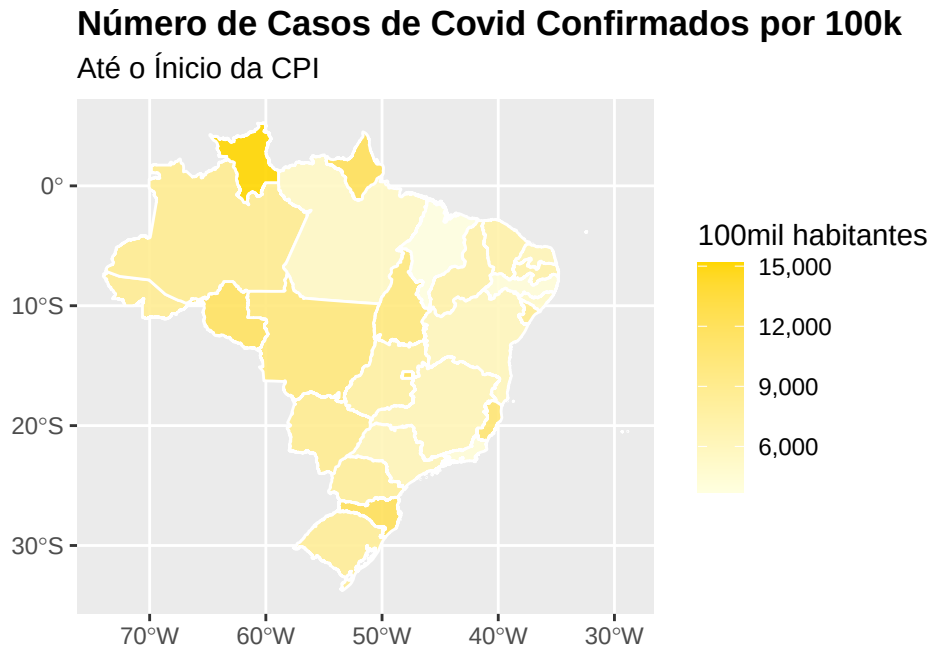


4.4 Dados Covid

Com a base construída neste trabalho, podemos analisar também os dados da covid e ver quais estados tiveram mais discursos contra a atuação do governo federal e quais sofreram mais com a pandemia em termos de casos confirmados por 100 mil habitantes e taxa de mortalidade. O primeiro mapa mostra a distribuição espacial por número de discursos contrários ao governo, com base no estado de origem dos senadores.



Vemos que os principais focos de discursos contrários à conduta federal são o Alagoas e o Amazonas. Não coincidentemente são os estados de origem de Renan Calheiros e Omar Aziz, que como vimos, são os dois senadores que tiveram mais discursos negativos. A seguir temos os mapas para o número de casos por 100 mil habitantes e a taxa de mortalidade de covid confirmados até o início da CPI.

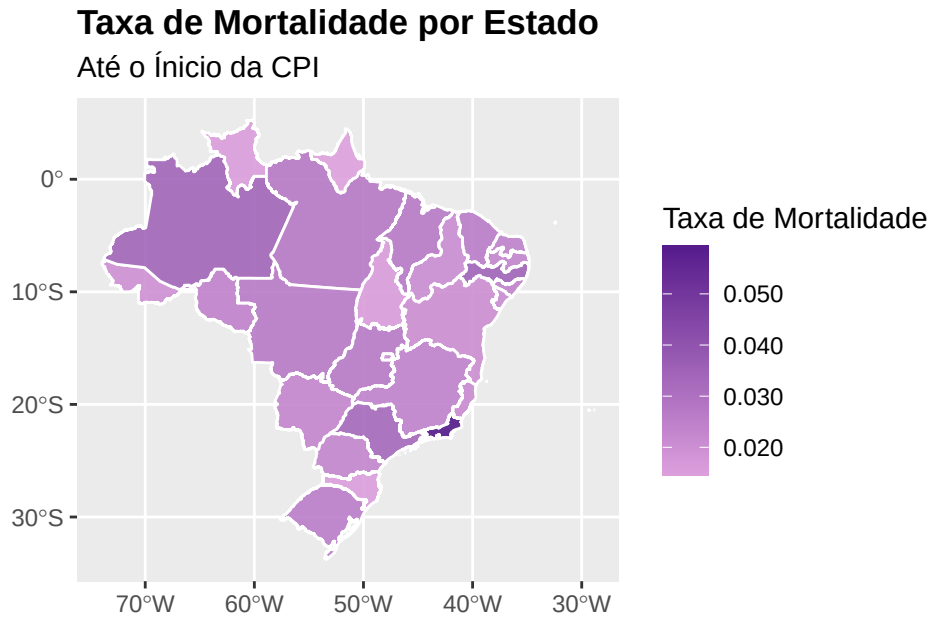


Pela análise dos 3 mapas em conjunto, é difícil afirmar se há uma correlação clara entre a intensidade da crise sanitária nos estados e o número de discursos contrários ao governo federal.

Disso algumas hipóteses surgem. Primeiramente, o estado exceção disso é o Amazonas, que no início do ano havia passado por uma enorme crise de saúde causada pelo coronavírus, faltando inclusive oxigênio para os pacientes. A crise lá certamente causou bastante comoção nacional e foi inclusive um dos motivadores para a instauração da CPI. Além disso, o presidente da comissão é crítico da atuação do governo nesse quesito.

Por outro lado, estados que se saíram bem no combate à pandemia utilizaram métodos diferentes dos quais o governo federal propagava. Então, é possível que mesmo com baixo número de casos e mortes relativos, eles sejam contrários à política da União justamente por verem o absurdo que seria se ela fosse levada como política pública de seus estados.

Entretanto, a mera visualização dos dados permite apenas especular. Análises econométricas mais robustas somadas à melhora do modelo de ML seriam necessárias para auferir resultados mais concretos.



5 Conclusão e Limitações

Infelizmente, por uma restrição temporal, não foi possível apresentar o trabalho e realizar todas as análises pensadas inicialmente. Algumas melhorias poderiam ser bem simples, mas acabaram não sendo priorizadas, como por exemplo a análise dos dados dos senadores.

Alguns nomes de senadores acabaram se perdendo. Seria necessário utilizar Regex para filtrar corretamente nomes como dos senadores Randolfe Rodrigues e Marcos Rogerio - personagens marcantes da CPI - ou então utilizar um dicionário manual de nomes para dar o match. Assim, a análise por gênero, raça e partido acabou ficando limitada e não tão representativa de toda a amostra, apesar de que espera-se que o viés não deve ser tão considerável.

Outro fator que poderia ter sido melhorado em relação ao modelo de ML é que a imensa maioria das falas não passavam informação alguma em relação ao que se estava tentando medir. Há também muitas interrupcoes nelas, e falas que nao são discursos de fato, mas palavras de ordem. Na base existiam colunas que indicavam questões de ordem e intervenções fora do microfone, por exemplo, mas elas acabavam filtrando muitos casos que gostaríamos de medir, e deixando passar muitos dos quais nós não gostaríamos. Portanto, optou-se por não utilizá-las como filtro.

Porém, uma medida que poderia ser tomada era tentar tratar melhor esses dados de forma a filtrar essas falas que não nos transmitem informações. Também poderia-se pensar em maneiras de juntar as partes dos discursos interrompidos transformando-os em uma fala coesa e contínua.

Outra forma de também aprimorar o modelo seria aumentando o número de observações da base de treino. Se tivéssemos um n=2000 ou n=3000, provavelmente teríamos um *fit* melhor, já que haveriam mais exemplos das classes para a máquina aprender os padrões corretamente.

Por fim, com o modelo bem especificado e outros métodos econométricos, como discutido da seção anterior, poderíamos auferir melhor a relação entre a crise sanitária nos estados e a posição dos senadores na CPI.