Universidade de São Paulo
Instituto de Matemática e Estatística

Centro de Estatística Aplicada

Relatório de Análise Estatística

RAE-CEA–24P16

RELATÓRIO DE ANÁLISE ESTATÍSTICA SOBRE O PROJETO:

"Multilateralismo, Regionalismo e Democracia: O uso das estratégias de shaming nas Missões de Observação Eleitoral da OEA"

Bruno Kinsch de Lara Campos
Daniel Monteiro Gallo
Rafael Bassi Stern

São Paulo, novembro de 2024

CENTRO DE ESTATÍSTICA APLICADA - CEA – USP

TÍTULO: Relatório de Análise Estatística sobre o Projeto: "Multilateralismo, Regionalismo e Democracia: O uso das estratégias de shaming nas Missões de Observação Eleitoral da OEA".

PESQUISADORES: Lucas Damasceno Pereira
Janina Onuki

ORIENTADOR: Prof. Rafael Bassi Stern

INSTITUIÇÃO: USP

FINALIDADE DO PROJETO: Doutorado

RESPONSÁVEIS PELA ANÁLISE: Bruno Kinsch de Lara Campos
Daniel Monteiro Gallo
Rafael Bassi Stern

REFERÊNCIA DESTE TRABALHO: CAMPOS, B.K.L.;GALLO, D.M.;STERN, R. S.

Relatório de análise estatística sobre o projeto: "Multilateralismo, Regionalismo e Democracia: O uso das estratégias de shaming nas Missões de Observação Eleitoral da OEA". São Paulo, IME-USP, 2024.

(RAE–CEA-24P16)

FICHA TÉCNICA

REFERÊNCIAS BIBLIOGRÁFICAS:

CORPORACIÓN LATINOBARÓMETRO: **Latinobarómetro: Opinión Pública Latinoamericana.** Disponível em: <https://www.latinobarometro.org> Acesso em: 16 de novembro de 2024

GARNETT, H.A.; JAMES, T.S.; CAAL-LAM, S. (2024): **Perceptions of Electoral Integrity (PEI-10.0)**. Disponível em: <https://dataverse.harvard.edu/dataverse/PEI> Acesso em: 16 de novembro de 2024

IZENMAN A.J. (2008) **Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning**. 1.ed. Springer. 733p.

PÉREZ, J.M.; RAJNGEWERC, Mariela; GIUDICI, J.C.; FURMAN, D.A.; LUQUE, Franco; ALEMANY, L.A.; MARTÍNEZ, M.V. (2021). **pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks.** Disponível em: <https://arxiv.org/abs/2106.09462> Acesso em: 29 de setembro de 2024

PROGRAMAS COMPUTACIONAIS UTILIZADOS:

Microsoft Word for Windows (versão 2016)

Microsoft Excel for Windows (versão 2016)

Python (versão 3.11.9)

R for Windows (versão 4.3.1)

RStudio for Windows (versão 6.1.524)

TÉCNICAS ESTATÍSTICAS UTILIZADAS

Análise Descritiva Multidimensional (03:020)

Análise de Associação e Dependência de Dados Quantitativos (06:010)

ÁREAS DE APLICAÇÃO

Resumo

Missões de observação eleitoral são missões internacionais que devem verificar a qualidade de processos eleitorais e emitir informes imparciais e transparentes relatando o que presenciaram. Todavia, dependendo da situação em que um país se encontra, um informe transparente que relata uma situação polêmica pode ser o gatilho para violência no país em observação. Nesse caso, existe a possibilidade de que as missões, a fim de não aumentar as tensões, usem estratégias de *shaming* (críticas públicas não vinculantes) nos informes.

Nesse contexto, foi proposto um estudo do *shaming* em diversos informes de observação eleitoral de diversos países da América Latina em relação à situação política dos países durante o período eleitoral. O *shaming* foi mensurado através da polaridade dos informes, enquanto a situação política foi analisada a partir de fatores como a presença de protestos pacíficos, protestos violentos, questionamento do resultado eleitoral, entre outras. Para entender como essas variáveis se relacionam, foi ajustado um modelo, usando técnicas Análise de Componentes Principais e regressão LASSO, que indica o quanto cada uma dessas variáveis impacta o *shaming*.

A partir da interpretação do modelo ajustado, conclui-se que variáveis como violência eleitoral, questionamento de resultados e preferência por regimes autoritários estão fortemente associadas ao fator *shaming*, enquanto elementos de estabilidade democrática contribuem para maior positividade nos informes. Esses resultados destacam a importância de compreender os fatores associados ao *shaming*, auxiliando na elaboração de estratégias de comunicação mais eficazes em missões de observação eleitoral

Sumário

1. Introdução

As eleições são uma parte essencial do processo democrático em diversos países, garantindo a legitimidade e a representatividade dos governantes. Para assegurar a integridade e a qualidade desses processos, missões de observação eleitoral são realizadas com o intuito de monitorar e relatar o andamento das eleições. Nesse contexto, as missões da Organização dos Estados Americanos (OEA) destacam-se como uma das principais formas de verificar a qualidade eleitoral na América Latina e no Caribe.

Ao fim das eleições, as missões de observação eleitoral da OEA emitem informes e notas para avaliar a qualidade do processo eleitoral. Um informe eleitoral da OEA é um documento oficial no qual a organização apresenta uma avaliação detalhada sobre a condução do processo eleitoral em um país-membro. Esses informes são baseados em dados coletados por observadores durante todo o ciclo eleitoral, desde a fase de preparação até a apuração dos votos. O objetivo é garantir que o processo eleitoral seja transparente, inclusivo e alinhado aos princípios democráticos internacionais. Os informes costumam incluir observações sobre aspectos técnicos da eleição, como o funcionamento das urnas, o acesso ao sufrágio e a neutralidade das autoridades eleitorais, além de destacar eventuais irregularidades ou fragilidades que possam comprometer a legitimidade do processo. Em alguns casos, os informes também apresentam recomendações para o aperfeiçoamento dos sistemas eleitorais nos países analisados, buscando promover o fortalecimento das instituições democráticas.

O presente estudo parte de um questionamento central: existe um viés político nos informes da OEA? Embora essas missões tenham o dever de relatar com precisão a qualidade dos processos eleitorais, é possível que as condições políticas internas de cada país influenciam a maneira como as críticas são formuladas. Um informe polêmico em meio a um cenário de instabilidade, por exemplo, pode agravar tensões sociais e até desencadear episódios de violência. Assim, a pesquisa busca analisar como as críticas públicas (*shaming*) são utilizadas.

2. Objetivos

O objetivo principal desta pesquisa é analisar os informes e notas de imprensa emitidos pelas missões de observação eleitoral da Organização dos Estados Americanos (OEA), com foco na utilização de estratégias de *shaming* (críticas públicas não vinculantes). O estudo visa identificar os fatores relevantes para a aplicação dessas críticas e avaliar as condições políticas, democráticas e eleitorais dos países observados. Além disso, o estudo propõe a utilização de uma variável indicativa de sinais de ruptura com o regime representativo.

3. Descrição do estudo

O estudo utiliza como fonte de dados os informes e notas de imprensa sobre as missões de observação eleitoral da OEA, complementados por dados pré-existentes sobre a qualidade das eleições, da observação eleitoral e da democracia dos países analisados, fornecidos pelos bancos *Perceptions of Electoral Integrity* (Pérez et al, 2021) e *Latinobarómetro* (Corporación Latinobarómetro, 2024). Os informes da OEA são divididos em dois tipos principais: eleitorais, que abordam o processo eleitoral de maneira técnica, e de imprensa, que são divulgados publicamente e têm maior impacto sobre a opinião pública. O foco principal desta pesquisa são os informes de imprensa, pois, ao serem disponibilizados publicamente, eles exercem uma maior influência na crítica pública.

Os informes estão disponíveis em quatro idiomas: inglês, francês, espanhol e português, e em dois tipos de eleições: presidenciais e legislativas. No entanto, este estudo considera apenas os informes em espanhol de eleições presidenciais. O idioma analisado compõe a maior parte dos documentos disponíveis, dado que a maioria dos países membros da OEA tem o espanhol como língua oficial, e analisar estes países é de maior relevância para entender a dinâmica política na região. Além disso, eleições presidenciais possuem maior impacto na população e são os principais cenários para golpes de estado, objeto de muito interesse neste estudo.

Utilizando o ano e o país de origem de cada informe como método para cruzar informações, foram adicionadas à base de informes variáveis externas obtidas das fontes descritas a seguir.

O *Electoral Integrity Project* é um projeto acadêmico independente, fundado em 2012, que tem como objetivo estudar as condições que levam a falhas em processos eleitorais e propor medidas para mitigá-las. Entre os produtos deste projeto destaca-se um banco de dados denominado *Perceptions of Electoral Integrity*, que avalia a qualidade das eleições ao redor do mundo com base na opinião de especialistas. Este banco reúne informações sobre eleições realizadas desde a fundação do projeto em diversos países, a versão utilizada no estudo é a *PEI10.0*. As variáveis extraídas estão detalhadas na próxima seção.

A *Corporación Latinobarómetro* é uma organização privada sem fins lucrativos, com sede em Santiago, Chile, que conduz periodicamente pesquisas de opinião pública nos países da América Latina. Por meio de questionários, são coletadas opiniões sobre diversas questões políticas e sociais. Dessa pesquisa, foram extraídas as proporções de respostas, por país e ano, à pergunta: “Com qual frase você mais concorda?”. As opções de resposta são: “A democracia é preferível a qualquer outro tipo de governo”, “Dentro de determinadas circunstâncias, um governo autoritário pode ser preferível a um democrático.”, “Para mim, não importa se temos um regime democrático ou não democrático.” e “Não sei”. Também foram coletadas as proporções daqueles que não responderam.

4. Descrição das variáveis

Variáveis que caracterizam a amostra:

- Informes: relatórios de imprensa escritos em língua espanhola sobre missões de observação eleitoral da OEA.
- País de origem do informe: Paraguai, El Salvador, Colômbia, Honduras, México, Peru, Brasil, Nicarágua, Equador, Bolívia, República Dominicana, Panamá, Suriname, Costa Rica, Guatemala, Guiana, Estados Unidos, Granada, Venezuela.

Variáveis relativas à análise textual e sentimental dos informes:

- Comprimento (≥ 0): número de palavras em um determinado documento.
- Diversidade Lexical (0 a 1): proporção de palavras únicas em um determinado informe.
- Polaridade (-1 a 1): carga emocional de um determinado informe. Valores negativos estão relacionados a sentimentos negativos, e valores positivos a sentimentos positivos.
- Entropia (≥ 0): medida de diversidade e dispersão de palavras. Valores mais altos indicam menor previsibilidade dos próximos termos de um texto.

Variáveis relativas ao banco *PEI10.0*:

- PEI (0 a 100): índice geral do banco *Perceptions of Electoral Integrity*. Mede a integridade eleitoral. Ou seja, a qualidade da eleição.
- Violence (1 a 5): nível da percepção de tratamentos violentos a eleitores.
- Protestspeace (1 a 5): nível da percepção de protestos pacíficos.
- Protestsviolent (1 a 5): nível da percepção de protestos violentos.
- Disputes (1 a 5): nível da percepção de disputas que foram resolvidas por meios legais.
- Results (0 a 100): índice aditivo e padronizado das variáveis Violence, Protestspeace, Protestsviolent, Disputes e Results .

Variáveis relativas ao banco *Latinobarómetro*:

- Dá no mesmo (0 a 1): proporção da população indiferente em relação ao tipo de governo.
- Democracia (0 a 1): proporção da população que prefere um governo democrático.
- Governo autoritário (0 a 1): proporção da população que prefere um governo autoritário.
- Não respondeu (0 a 1): proporção da população que não respondeu a questão.
- Não sabe (0 a 1): proporção da população que não sabe responder a questão.

5. Análise descritiva

Apresentaremos, a seguir, a análise descritiva dos dados. Esta etapa visa caracterizar e resumir os principais aspectos observados, oferecendo uma visão geral inicial sobre as distribuições das variáveis e destacando comportamentos que poderão ser explorados em análises subsequentes.

Para facilitar a análise, os informes passaram por um pré-processamento de textos. Esse processo envolveu várias etapas, como a conversão de letras maiúsculas em minúsculas, a remoção de pontuações e números, e a exclusão de palavras irrelevantes (*stopwords*) utilizando a biblioteca *nltk* (Natural Language Toolkit) do Python (Pérez et al, 2021). Adicionalmente, foi gerado um segundo conjunto de informes com um tratamento mais específico, do qual removemos manualmente *stopwords* adicionais. Essas palavras, apesar de não serem eliminadas por métodos automáticos, foram consideradas irrelevantes para a análise devido ao contexto político dos informes, tornando a discriminação das cargas emocionais mais precisa. As *stopwords* adicionadas manualmente estão explícitas na Tabela A.3. Os plurais também foram incluídos.

As análises foram realizadas utilizando ambos os conjuntos de informes: o que contém apenas as *stopwords* removidas automaticamente pela biblioteca *nltk* e o conjunto com as *stopwords* adicionais removidas manualmente. A comparação entre esses dois conjuntos permite avaliar o impacto da exclusão dessas palavras adicionais.

5.1 Análise quantitativa dos textos

Os primeiros gráficos gerados na análise foram "nuvens de palavras", uma visualização utilizada para resumir a frequência de termos em um conjunto de textos. Nesse tipo de gráfico, as palavras que aparecem com maior frequência nos documentos são exibidas em tamanhos proporcionais à sua ocorrência.

Na Figura B.1, aplicamos essa técnica ao conjunto de informes sem a remoção de *stopwords*, muitas palavras aparecem com tamanhos semelhantes, como "observación electoral", "proceso electoral" e "misión observación". Esses termos, apesar de aparecerem com alta frequência, são esperados neste tipo de

documento, uma vez que são pontos centrais dos temas abordados. A alta presença destes termos reduz o destaque de palavras que poderiam ser mais relevantes para a análise, já que estes não são informativos sobre a polaridade dos textos.

No conjunto de informes para os quais foram removidas as *stopwords* adicionais, a nuvem de palavras apresenta termos mais relevantes (Figura B.2), como "sociedad civil", "supremo", "desarrollo" e "ciudadanía". Esses termos estão mais relacionados a questões políticas e sociais específicas, sendo úteis para a compreensão do conteúdo e do sentimento dos informes.

Em seguida, foram construídos gráficos de barras, como uma outra forma de representar a frequência das palavras nos dois conjuntos de informes: na Figura B.3, sem a remoção das *stopwords*, destacam-se as palavras "electoral", "misión" e "oea", que dominam fortemente a frequência. Esses termos estão relacionados ao tema dos informes, como o processo eleitoral e às atividades da OEA. Embora frequentes, não são necessariamente úteis para discriminar aspectos mais específicos dos textos.

Com a remoção das *stopwords*, podemos observar na Figura B.4, que o gráfico de barras apresenta uma distribuição mais homogênea, destacando palavras que parecem diferenciar tópicos mais específicos dos informes, como "financiamento", "ciudadanía", "mujeres" e "guatemala". Essas palavras fornecem uma visão mais precisa dos temas discutidos, o que sugere que a eliminação das *stopwords* adicionais ajuda a expor tópicos de maior relevância.

Foram construídos também os gráficos de coocorrência de bigramas, que mostram a frequência com que pares de palavras (bigramas) aparecem juntos nos textos. Esses gráficos são úteis para identificar termos que frequentemente ocorrem em sequência, revelando padrões de associação entre palavras dentro dos informes.

No gráfico de coocorrência dos bigramas para o conjunto que inclui as *stopwords*, apresentado na Figura B.5, as combinações mais destacadas são "misión observación" e "observación electoral". Esses bigramas estão fortemente relacionados ao tema central dos informes, que frequentemente descrevem

missões de observação eleitoral, mas também não oferecem grande valor discriminativo para a análise mais detalhada dos textos.

No conjunto sem as *stopwords* adicionais, apresentado na Figura B.6, o gráfico de coocorrência apresenta bigramas mais informativos, como "costa rica" aparecendo com "juntas receptoras" e "medios comunicación", assim como "sociedad civil" aparecendo com "medios comunicación". Esses bigramas sugerem uma relação entre temas específicos abordados nos informes.

Os gráficos das Figuras B.7 e B.8 são histogramas que ilustram a diversidade lexical dos informes. A diversidade lexical é uma métrica que mede a variedade de palavras diferentes em um texto, sendo útil para avaliar o quão complexo é o vocabulário de um documento. Esta medida é uma proporção, em que o valor 1 indica que toda palavra no texto aparece uma única vez.

O histograma da Figura B.7 inclui as *stopwords* e sua distribuição é relativamente simétrica, com a maior concentração de valores variando entre 0,73 e 0,78. Isso sugere que, apesar de um vocabulário diverso, as stopwords estão contribuindo para uma menor distinção entre os textos. Quando removemos as *stopwords* adicionais, o histograma se desloca para valores mais altos, com a maioria dos informes apresentando uma diversidade lexical entre 0,86 e 0,9, como visto na Figura B.8. Esse aumento reflete uma maior variedade de palavras significativas, já que as palavras mais comuns e repetitivas foram excluídas. Em ambos os conjuntos, também observamos a presença de alguns informes com diversidade lexical bastante baixa, mas esses casos são minoria.

Os gráficos das Figuras B.9 e B.10 tratam da entropia dos informes. A entropia, em um contexto de análise de textos, mede o grau de imprevisibilidade ou desordem das palavras. Assim como a diversidade lexical, a entropia é uma métrica que avalia a riqueza de termos em um documento, mas enquanto a diversidade lexical foca na quantidade de palavras diferentes, a entropia leva em consideração também a frequência e a distribuição dessas palavras. Entropias altas indicam que a distribuição das palavras em um texto é mais uniforme, ou seja, não há uma grande concentração de frequência em algumas palavras específicas. Em outras palavras, em um texto com alta entropia, as palavras estão distribuídas de forma mais equilibrada, sem que algumas apareçam de forma

predominante.

Na Figura B.9, estão inclusas as *stopwords*. Nota-se uma leve assimetria à direita, com a maior concentração de valores de entropia variando entre 6,3 e 7,3. Quando as *stopwords* adicionais são removidas, observa-se na Figura B.10 que o histograma mostra uma distribuição semelhante, com uma leve assimetria à direita e concentração de valores entre 6,2 e 7,4. A distribuição é muito próxima da primeira, sugerindo que as *stopwords* não são tão impactantes na entropia dos textos.

As medidas-resumo relativas à diversidade lexical e à entropia dos informes, podem ser verificadas nas Tabelas A.1 e A.2.

5.2 Análise da polaridade dos textos

Para a análise da polaridade dos informes, foi utilizada a biblioteca *pysentimiento* (Pérez et al., 2021). Dentro dessa biblioteca, existe uma função que classifica os textos e retorna as probabilidades de um texto ser classificado como negativo, neutro e positivo.

Com base nessas probabilidades, foi desenvolvida uma nova medida de polaridade que varia de -1 a 1. Nessa escala, -1 representa uma polaridade extrema negativa, 0 indica uma polaridade neutra e 1 denota uma polaridade extrema positiva. Essa abordagem permite uma interpretação mais intuitiva do sentimento geral expressado nos textos, facilitando a comparação entre diferentes informes. Além disso, foram calculadas as medidas-resumo para esta medida, podendo ser verificadas nas Tabelas A.1 e A.2.

Nas Figuras B.11 e B.12, temos os histogramas das polaridades dos informes na escala definida anteriormente. Na Figura B.11, que inclui as *stopwords*, observa-se que a grande maioria dos informes apresenta uma polaridade variando entre -0,04 e 0,2. Essa concentração sugere que a maioria dos textos possui um tom levemente positivo ou neutro. No entanto, é importante notar que também há a presença de alguns informes com polaridade negativa alta, variando entre -0,17 e -0,79. Esses casos extremos indicam que, embora a maioria dos informes transmita um sentimento relativamente neutro ou positivo, existem alguns textos que podem expressar críticas severas.

Quando analisamos o gráfico sem as *stopwords*, apresentado na Figura B.12, a distribuição não apresenta grandes alterações. A concentração dos dados se encontra entre -0,01 e 0,24, o que reforça a ideia de que a maior parte dos informes mantém um tom levemente positivo ou neutro. No entanto, a presença de informes com polaridade negativa extrema continua nesta distribuição. Isto sugere que a remoção das *stopwords* não impactou significativamente a interpretação do sentimento geral.

Os gráficos das Figuras B.13 e B.14 ilustram as palavras mais frequentes nos informes positivos e nos informes negativos, respectivamente. Para a construção destes gráficos, foram considerados os 50 informes mais extremos positivamente e os 50 informes mais extremos negativamente. Esses gráficos permitem identificar quais termos aparecem com maior destaque em textos que transmitem sentimentos mais intensos.

Nos informes com polaridade mais positiva, destacam-se termos como "financiamento", "mujeres", "costa rica", "felicita" e "colombia". Essas palavras podem estar correlacionadas com o tom otimista dos informes. Palavras como "felicita" possuem polaridade positiva por si só, enquanto termos como "costa rica", "colombia" e "mujeres" podem estar associados a assuntos positivos específicos de cada informe.

Já nos informes com polaridade mais negativa, podemos destacar os termos "guatemala", "nicaragua", "escrutinio", "violencia" e "preocupación". Esses termos indicam questões delicadas, como violência e problemas no processo eleitoral (refletido pela palavra "escrutinio"). Além disso, a menção a países como Guatemala e Nicarágua, pode estar relacionada a situações específicas de crise ou crítica ao governo destes países.

Por fim, nas Figuras B.15 a B.17 foram construídos diagramas de dispersão para verificar a correlação entre a polaridade dos textos e outras variáveis analisadas. Os três gráficos associam a polaridade com o comprimento do texto, com a diversidade lexical e com a entropia, respectivamente. Em todos os casos, a correlação encontrada foi bem baixa, próxima de 0. Isso sugere que a carga emocional de um texto não está relacionada diretamente ao seu comprimento, à sua diversidade de vocabulário ou à sua imprevisibilidade.

Em seguida, na Figura B.18, foi criado um *box plot* para analisar a distribuição da polaridade dos textos por país de origem dos informes. Alguns países, como Brasil, Costa Rica, Paraguai e Granada, destacam-se por apresentarem uma distribuição de polaridade com tendências mais positivas. Por outro lado, países como Peru e Guatemala tendem a apresentar uma polaridade mais negativa. No entanto, é notável que *outliers* negativos estão bem distribuídos entre os diferentes países, sugerindo que, independentemente do contexto geral de cada país, existem casos de informes mais críticos.

Foi analisada também, na Figura B.19, a correlação entre a entropia e a diversidade lexical dos textos, que apresentou uma correlação negativa significativa. Isso indica que, quanto maior a diversidade lexical, menor tende a ser a entropia. Esse resultado pode ser explicado pelo fato de que textos com alta diversidade lexical utilizam muitas palavras diferentes, enquanto aqueles com alta entropia tendem a ter uma distribuição mais uniforme das palavras, sem um foco predominante em termos específicos.

Dada essa correlação significativa entre entropia e diversidade lexical, considera-se que, ao criar uma medida baseada em polaridade, entropia e diversidade lexical, o ideal seria optar por uma dessas duas últimas variáveis, para evitar problemas de multicolinearidade. A polaridade não apresenta correlação significativa com entropia ou diversidade lexical, o que não gera problemas nesse aspecto.

Apesar disso, optou-se por utilizar apenas a polaridade do informe como representação do aspecto de *shaming* que ele transmite, uma vez que o fator sentimental capturado por essa variável foi considerado suficiente para os objetivos do estudo. Além disso, ao limitar a análise a uma única variável, preserva-se tanto a simplicidade quanto a interpretabilidade da medida, eliminando a necessidade de justificativas adicionais sobre as ponderações que seriam aplicadas na criação de um índice.

Assim, definimos a polaridade como a variável resposta do estudo.

5.3 Análise das covariáveis

Analisaremos, a seguir, as covariáveis (variáveis independentes) provenientes dos outros bancos de dados apresentados, que desejamos associar à polaridade dos informes, portanto, o foco principal desta seção é descrever graficamente esta associação. A partir desta etapa da análise, passamos a utilizar a linguagem R, uma escolha mais comum para este tipo de estudo, considerando que o uso da *pysentimiento* não é mais necessário.

Começando pelas variáveis quantitativas, faremos a análise por meio de gráficos de dispersão e do cálculo do coeficiente de correlação linear de Pearson, que mensura a direção e a intensidade da associação. Essa abordagem já foi utilizada na seção anterior. As Figuras B.20 e B.21 apresentam os gráficos de dispersão com os coeficientes de correlação e as linhas de tendência para as covariáveis PEI e Results. Observa-se que a correlação para PEI é bem próxima de 0. Isso também é evidenciado pela linha de tendência, quase paralela ao eixo horizontal, e pelo comportamento dos pontos, que formam uma nuvem dispersa, sem um padrão claro da polaridade à medida que o valor de PEI aumenta. Para Results, a associação é um pouco mais definida: com uma correlação de 0,212, é possível identificar a inclinação da linha de tendência, indicando uma tendência de aumento na polaridade conforme Results aumenta.

Os gráficos de dispersão das proporções de respostas para a questão do *Latinobarómetro* estão representados nas Figuras B.22 a B.25. A partir desses gráficos, nota-se que as intensidades de associação da polaridade com cada resposta possível são semelhantes, com exceção dos que não sabiam ou não responderam, cuja correlação é próxima de 0. As proporções com correlação negativa relativamente significativa em relação à polaridade são as daqueles que preferem governos autoritários em determinadas circunstâncias e dos que não têm preferência entre regimes democráticos ou autoritários. Isso gera indícios de que, quando essas proporções aumentam, a carga emocional negativa dos informes tende a crescer. Por outro lado, há indícios de que a carga emocional positiva é maior para proporções mais elevadas de respondentes que preferem regimes democráticos a qualquer outro tipo de governo, visto que essa relação apresenta correlação positiva.

Para as variáveis categóricas, que neste caso, são ordinais com categorias de 1 a 5, a representação mais adequada é por meio de *box plots*. Esses gráficos são utilizados para analisar a distribuição de uma variável resposta quantitativa em diferentes níveis de uma variável categórica. Além de ilustrar a distribuição da variável de interesse por meio do intervalo entre o primeiro e o terceiro quartil, os *box plots* também evidenciam a presença de *outliers* (valores atípicos), caso existam.

As Figuras B.26 a B.30 apresentam os *box plots* para cada covariável categórica. Observa-se que os níveis 2 e 3 da variável Violence apresentam alta heterogeneidade, o que é evidenciado pelos amplos intervalos interquartis, isso dificulta a identificação de uma diferença significativa entre um nível e outro. Por outro lado, nota-se que a polaridade no nível 4 de Violence tende a se concentrar em valores menores em comparação ao nível 1. Isso sugere a hipótese de que um aumento na violência eleitoral pode estar associado a uma carga emocional mais negativa nos informes.

Aplicando a mesma análise às variáveis Challenged e Protestspeace, levantamos hipóteses semelhantes: à medida que os níveis dessas variáveis aumentam, a polaridade parece diminuir. No entanto, para as variáveis Protestsviolent e Disputes, essa associação não é tão evidente apenas pela análise visual. Além disso, observa-se a presença de diversos *outliers* com polaridade muito baixa em todos os gráficos, indicados pelos pontos destacados. Esses *outliers* representam os informes mais críticos, caracterizados por uma polaridade extremamente baixa e maior intensidade no fator de *shaming*.

A partir dessas análises preliminares, ainda não é possível realizar conclusões definitivas. Contudo, essas observações fornecem uma ideia inicial do comportamento dos dados e ajudam a direcionar a análise inferencial para os aspectos mais notáveis identificados.

6. Análise inferencial

A partir desta seção, inicia-se a análise inferencial do projeto, cujo objetivo é construir um modelo preditivo para a polaridade dos informes com base nos

valores das covariáveis. O foco principal é avaliar o efeito de cada covariável na variável resposta, analisando os coeficientes estimados do modelo para entender como a polaridade varia, em qual direção e com que intensidade, conforme os valores das covariáveis também variam.

6.1 Análise de Componentes Principais (ACP)

Antes de ajustar um modelo, foi realizada uma análise de componentes principais (ACP) nos dados. Essa é uma técnica estatística não supervisionada utilizada para reduzir a dimensionalidade da base de dados, ou seja, diminuir a quantidade de variáveis (Izenman, 2008). No contexto deste projeto, a técnica foi aplicada às covariáveis, com exceção da PEI, que representa o índice geral de qualidade da democracia e que será analisada de forma isolada. Além disso, foram excluídas as colunas relativas às proporções das respostas “Não sei” e de não-respostas na questão do *Latinobarómetro*. A razão para isso é que a soma das proporções de todas as respostas é igual a 1, o que introduz um problema de multicolinearidade que poderia comprometer os resultados da ACP. Neste caso, os efeitos das demais respostas serão avaliados tomando o grupo das respostas “Não sei” e de não-respostas como referência.

A aplicação dessa técnica foi motivada, principalmente, por dois fatores. Primeiro, como observado na análise descritiva, as covariáveis não apresentaram uma associação clara com a variável resposta quando analisadas individualmente, mas, em conjunto, podem ter uma influência significativa. Segundo, algumas covariáveis possuem dependência entre si, como já mencionado no caso das proporções das respostas da questão do *Latinobarómetro*, e dos índices PEI e Results, que dependem de outras variáveis do mesmo banco de dados.

A ACP reduz a quantidade de covariáveis ao criar novas variáveis, chamadas de componentes principais, que são combinações ponderadas das variáveis originais. Esses componentes são construídos de forma a capturar, idealmente, a maior parte da variância gerada pelas variáveis iniciais, permitindo reduzir a dimensionalidade sem perda significativa de informação e lidando com as redundâncias presentes nas variáveis originais. Para isso, é necessário definir a

perda máxima tolerável de variância explicada ao reduzir as variáveis originais aos componentes principais. Um critério muito utilizado na literatura é tolerar até 10% de perda, ou seja, escolher uma quantidade de componentes principais que expliquem pelo menos 90% da variância total.

Utilizando a função nativa do R *prcomp,* aplicamos a análise de componentes principais. Na Figura B.31, apresentamos o *scree plot*, um gráfico que ilustra a variância explicada acumulada à medida que mais componentes principais são adicionados, ordenados decrescentemente pela quantidade de variância explicada individualmente. Observa-se que, com 9 componentes, é possível explicar 100% da variância, que é o número de variáveis utilizadas no PCA. No entanto, ao analisar o gráfico, nota-se que com apenas 4 componentes principais já é possível explicar um pouco mais de 90% da variância. Assim, optou-se por reduzir as covariáveis originais a 4 componentes principais.

Na Tabela A.4 são apresentadas as ponderações aplicadas às variáveis originais para a construção de cada componente principal, de CP1 (PC1) a CP4 (PC4). O produto final dessa análise foi a redução de 10 covariáveis para apenas 5: os 4 componentes principais construídos e a variável PEI. Essa redução permite ajustar o modelo preditivo sem preocupação com problemas de multicolinearidade. Além disso, ao utilizar um número menor de covariáveis, reduzimos a quantidade de coeficientes a serem estimados, o que impacta positivamente na variância das previsões geradas pelo modelo. Em outras palavras, o erro preditivo em novos conjuntos de dados tende a ser menor.

6.2 Regressão LASSO

O primeiro modelo ajustado para predizer a polaridade foi uma regressão LASSO, uma técnica semelhante à regressão linear convencional, pois busca estimar uma equação que relacione as covariáveis à variável resposta. No entanto, diferentemente da regressão linear, o LASSO (Izenman, 2008) aplica uma penalização aos coeficientes das covariáveis.

Na regressão linear convencional, os coeficientes são estimados minimizando o erro preditivo (função objetivo). Contudo, incluir muitas covariáveis

tende a reduzir o erro preditivo na base utilizada para ajuste, mas aumenta o erro em novos conjuntos de dados. Esse comportamento é conhecido como *overfitting* e compromete o ajuste do modelo. Para evitar isso, o LASSO penaliza a função objetivo com a soma dos valores absolutos dos coeficientes estimados, o que induz alguns coeficientes a serem zerados, isto é, algumas covariáveis não são utilizadas na equação preditiva final. Este procedimento pode ser visto como uma seleção das covariáveis mais importantes.

A função objetivo do LASSO depende de um parâmetro λ, que é um peso atribuído à intensidade da penalização. Valores diferentes de λ afetam os coeficientes e, consequentemente, o erro preditivo. Para encontrar o valor ideal de λ, utilizou-se validação cruzada. Nesse processo, a base foi dividida em 10 subamostras, e o modelo foi ajustado iterativamente, utilizando uma das subamostras para cálculo do erro preditivo e o restante para o ajuste de modelo. A cada iteração, a subamostra usada para o cálculo do erro é trocada, e um valor diferente para o erro preditivo é obtido. A média desses erros fornece o erro preditivo geral da validação cruzada que pode ser utilizado para avaliar a performance do modelo. Na Figura B.32, temos um gráfico em escala logarítmica que mostra o erro quadrático médio da validação cruzada para diferentes valores de λ. O valor que minimizou o erro foi 0,0091. Este gráfico foi obtido por meio das funções *glmnet,* da biblioteca de mesmo nome, para o ajuste da regressão LASSO, e da função *cv.glmnet*, da mesma biblioteca, para a realização da validação cruzada.

Ajustando o modelo com λ igual a 0,0091, obtemos os coeficientes estimados expressos na Tabela A.5. Nota-se que os coeficientes relativos ao PC2, PC4 e PEI foram zerados, ou seja, essas variáveis não foram significativas para a predição. Assim, a equação preditiva final é:

$$\hat{y} = 0{,}18 - 0{,}04 \cdot PC1 + 0{,}03 \cdot PC3,$$

em que $\hat{y}$ representa a polaridade esperada.

Para interpretar os resultados, analisamos a composição dos componentes principais (Tabela A.4). O coeficiente negativo para o PC1 indica que aumentos em variáveis com ponderações positivas como Violence, Challenged, Protestspeace, Protestsviolent, Dá no mesmo e Governo autoritário estão associados a uma

redução na polaridade, refletindo uma carga emocional mais negativa. Por outro lado, aumentos em variáveis como Disputes e Democracia estão associados a um aumento na polaridade, indicando uma carga emocional mais positiva.

É importante ressaltar também que os valores absolutos destas ponderações estão relacionados à intensidade do efeito das covariáveis sobre a polaridade. Portanto, verifica-se que para o PC1, o efeito da covariável Governo autoritário não é tão intenso quanto o efeito das demais covariáveis, pois sua ponderação é relativamente baixa (0,10).

No caso do PC3, o coeficiente positivo sugere que aumentos em covariáveis com ponderações positivas estão associados a aumentos na polaridade esperada. Sabendo disso, nota-se que algumas covariáveis têm efeitos contrários aos observados para o PC1. Isto é verificado para os efeitos de Protestsviolent e Results, que possuem intensidade semelhante em PC1 mas em direção contrária, o que torna questionável a significância destas covariáveis. Esta contradição também é observada em outras covariáveis, mas esses efeitos são menos relevantes devido aos valores absolutos significativamente menores das ponderações.

Todavia, os valores dos coeficientes de Dá no mesmo, Democracia e Governo Autoritário são grandes dentro do PC3, especialmente esse último. Assim podemos inferir que o PC3 seja mais especializado do que o PC1. Enquanto o PC1 parece uma mistura geral dos fatores para indicar a aderência às regras do sistema democratico, o PC3 foca na estrutura formal (Democracia- Autoritário) e no impacto de protestos violentos.

Podemos inclusive interpretar o PC1 como uma representação de uma situação política menos polarizada, onde diversos fatores importam para o *shaming* do relatório. E o PC3 representa uma situação mais polarizada, onde os fatores mais relevantes para o shaming do relatório são as variáveis Democracia, Governo Autoritário, Dá no mesmo, Results e Protestviolent. Ou seja, o quanto a população está aderindo ao sistema democrático, se os resultados da eleição foram questionados e se houve protestos violentos.

Dessa forma, talvez seja melhor nos referirmos ao PC1 como “USUAL” e ao PC3 como “POLARIZADO”, assim indicando o efeito e o peso desses fatores em

suas respectivas situações e facilitando a interpretação. Assim podemos reescrever a equação anterior de uma forma mais interpretável:

$$\hat{y} = 0{,}18 - 0{,}04 \cdot \text{USUAL} + 0{,}03 \cdot \text{POLARIZADO},$$

Consolidando as conclusões, espera-se que aumentos em variáveis como violência eleitoral (Violence), questionamento dos resultados eleitorais (Challenged), protestos pacíficos (Protestspeace), a indiferença em relação ao regime governamental (Dá no mesmo) e preferência por governos autoritários em determinadas circunstâncias (Governo autoritário) diminuam significativamente a polaridade, o que revela a significância destes aspectos para o fator *shaming*. Já variáveis como o índice de disputas resolvidas legalmente (Disputes) e preferência por regimes democráticos (Democracia) contribuem positivamente para a polaridade.

7. Conclusão

Conclui-se que o uso de estratégias de *shaming* nos informes de missões de observação eleitoral está fortemente associado a variáveis relacionadas à instabilidade política, como violência eleitoral, questionamento de resultados e preferências autoritárias, enquanto contextos de maior estabilidade democrática promovem informes mais positivos. Esses resultados indicam que o tom dos informes reflete a complexidade das situações políticas locais, destacando a importância de ajustar a comunicação para evitar a escalada de tensões nos países observados.

APÊNDICE A

Tabelas

Tabela A.1 *Medidas-resumo dos informes com as stopwords adicionais*

Medida	Comprimento	Diversidade Lexical	Entropia	Polaridade
Média	302,51	0,76	7,14	0,09
Desvio Padrão	416,97	0,07	0,82	0,16
Mínimo	15	0,45	3,77	-0,79
1º Quartil	140	0,74	6,57	0,05
Mediana	190	0,77	7,00	0,11
3º Quartil	287	0,80	7,53	0,16
Máximo	4516	0,93	10,01	0,45

Tabela A.2 *Medidas-resumo dos informes sem as stopwords adicionais*

Medida	Comprimento	Diversidade lexical	Entropia	Polaridade
Média	223,34	0,87	6,94	0,09
Desvio padrão	326,15	0,07	0,96	0,19
Mínimo	6	0,55	2,59	-0,78
1º Quartil	94	0,84	6,31	0,05
Mediana	135	0,87	6,80	0,11
3º Quartil	211	0,91	7,39	0,19
Máximo	3200	1,00	10,47	0,50

Tabela A.3 *Stopwords adicionadas manualmente na etapa de pré-processamento de texto*

Stopwords
oea, dijo, dia, organización, estados, americanos, general, america, ambassador, secretario, reunión, permanente, americas, consejo, miembros, país, hoy, gmt, electoral, misión, proceso, observación, elección, enero, febrero, marzo, abril, mayo, junio, julio, agosto, septiembre, octubre, noviembre, diciembre, partido, político, votación, participación, autoridade, resultado, parte, asimismo, sistema, jornada, jefe, observó, manera, voto, política, informe, candidato, observador, mesa, tribunal, información, nacional, comicio, toda, segunda, primera, campaña

Tabela A.4 *Ponderações das covariáveis na composição dos componentes principais.*

	PC1	PC2	PC3	PC4	PC5
Violence	0,30	-0,49	0,07	-0,25	-0,76
Challenged	0,38	0,27	0,03	0,31	0,01
Protestspeace	0,36	0,33	0,04	0,45	-0,28
Protestsviolent	0,34	0,02	0,40	-0,46	0,30
Disputes	-0,39	-0,04	-0,07	0,35	-0,22
Results	-0,39	-0,25	-0,23	-0,06	0,05
Dá no mesmo	0,31	-0,48	-0,29	0,36	0,21
Democracia	-0,33	0,27	0,45	0,02	-0,33
Governo autoritário	0,10	0,45	-0,70	-0,41	-0,21

Tabela A.5 *Coeficientes estimados para a regressão LASSO.*

Variável	Estimativa
Intercepto	0,18
PEI	.
PC1	-0,04
PC2	.
PC3	0,03
PC4	.

APÊNDICE B

Figuras

Figura B.1 *Nuvem de palavras total dos informes*

Figura B.2 *Nuvem de palavras dos informes sem as stopwords adicionadas*

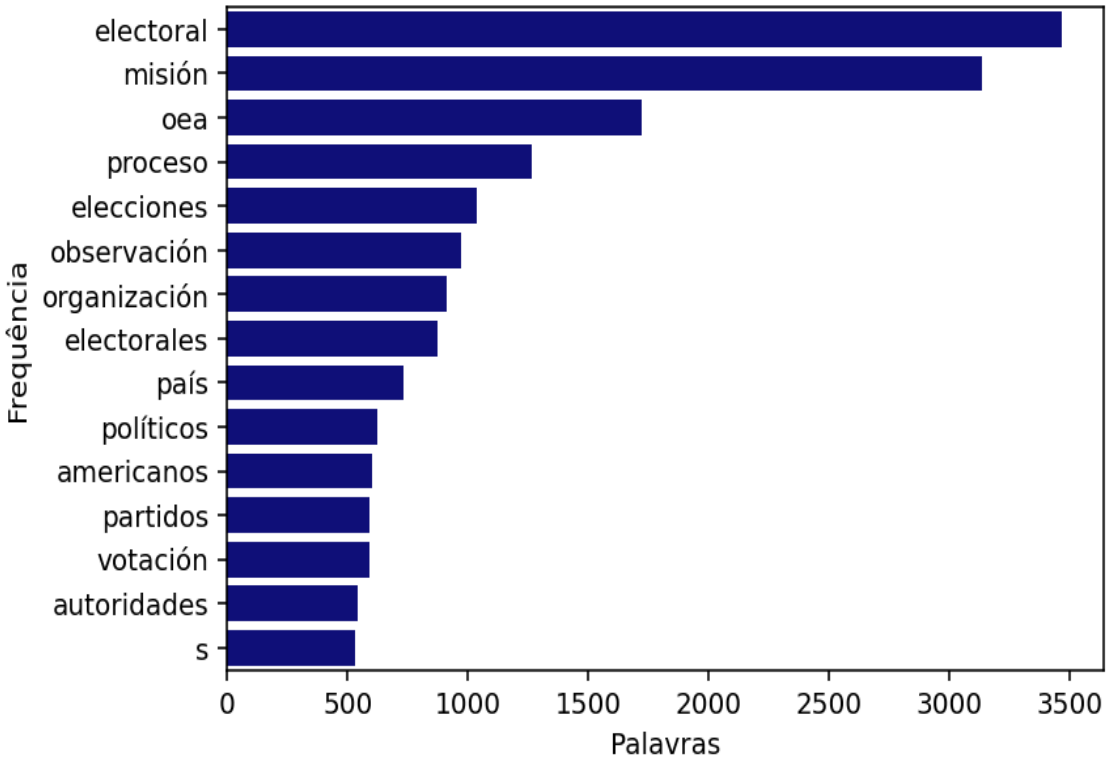

Figura B.3 *Palavras mais frequentes nos informes com as stopwords adicionadas*

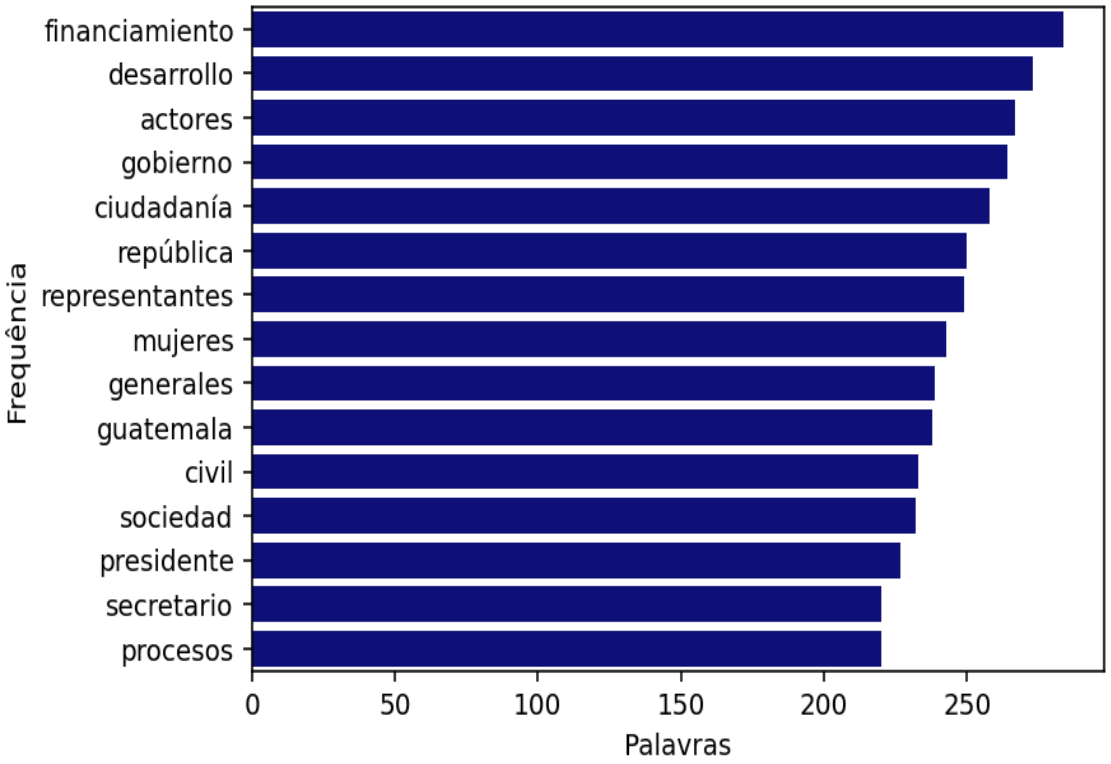

Figura B.4 *Palavras mais frequentes nos informes sem as stopwords adicionadas*

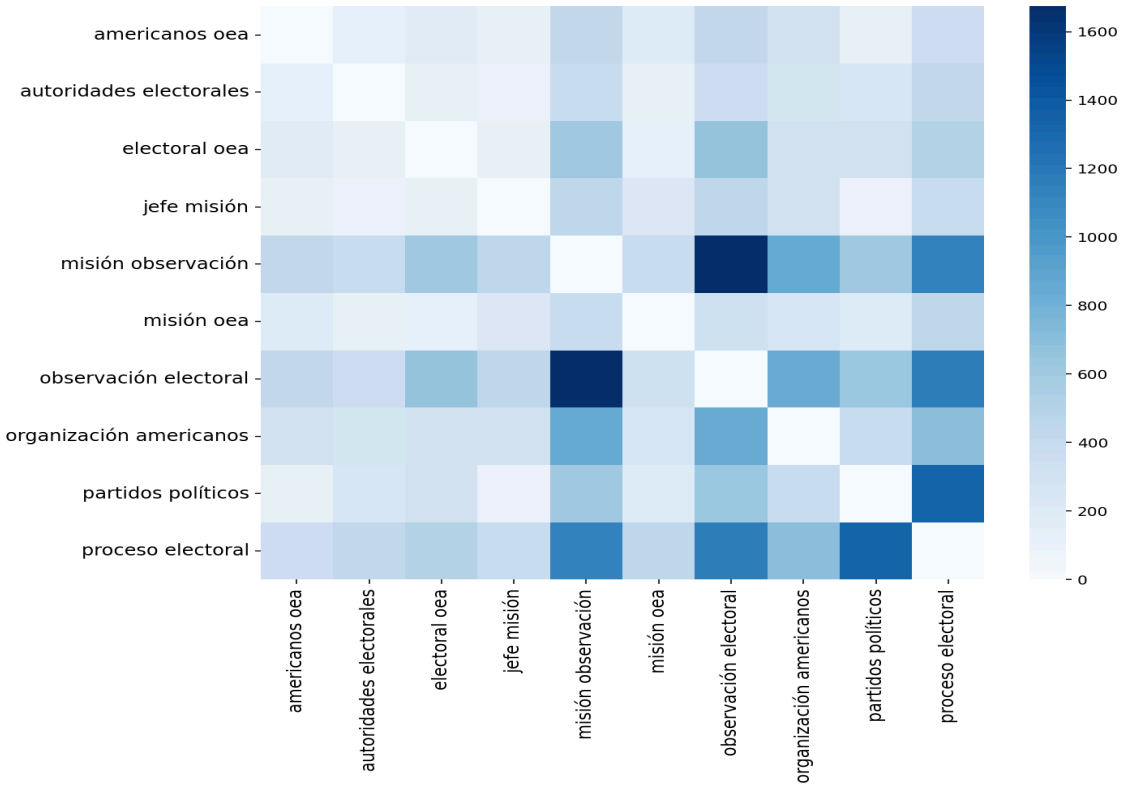

Figura B.5 *Coocorrência de bigramas nos informes com as stopwords adicionadas*

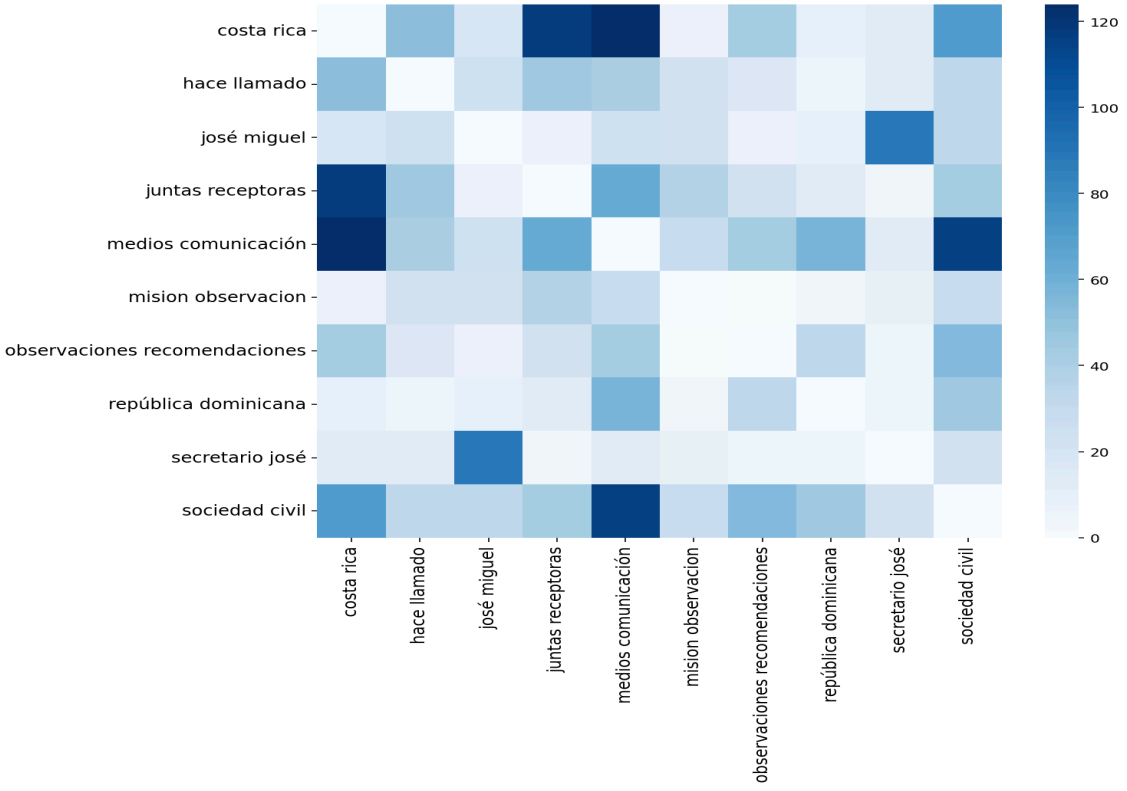

Figura B.6 *Coocorrência de bigramas nos informes sem as stopwords adicionadas*

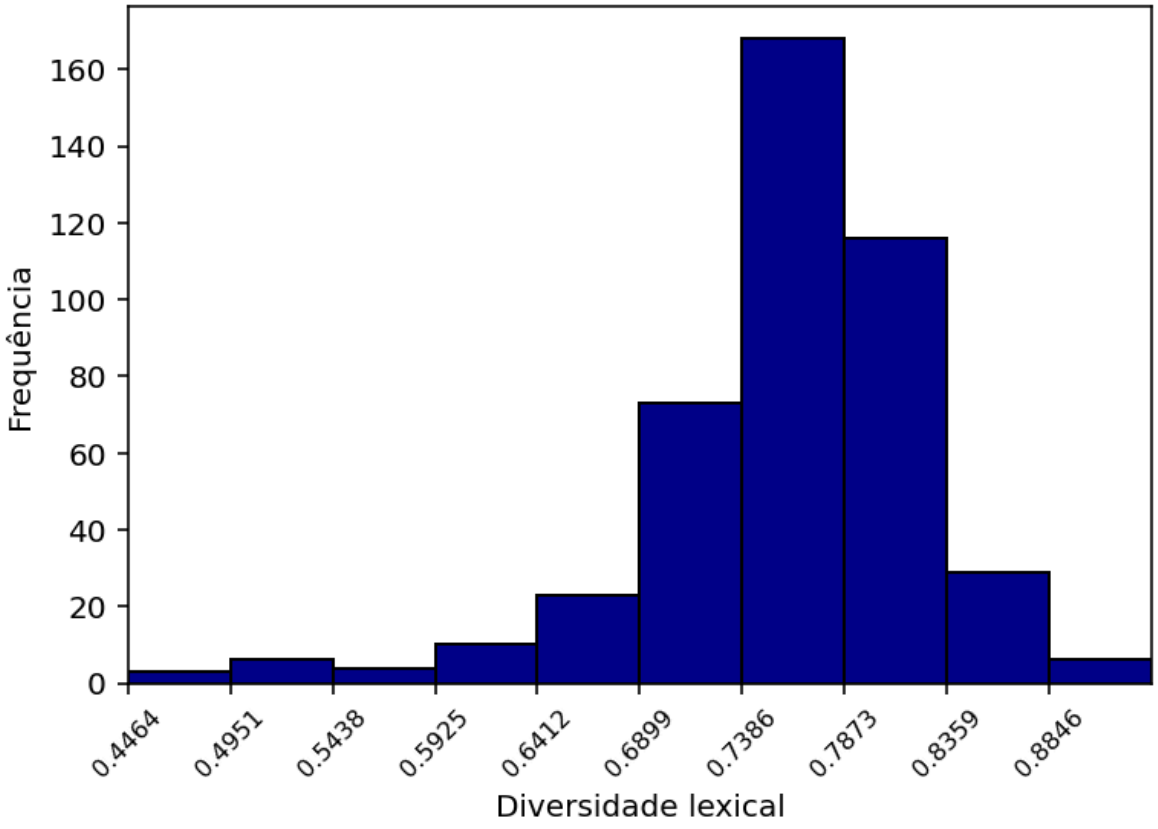

Figura B.7 *Distribuição da diversidade lexical com as stopwords adicionadas*

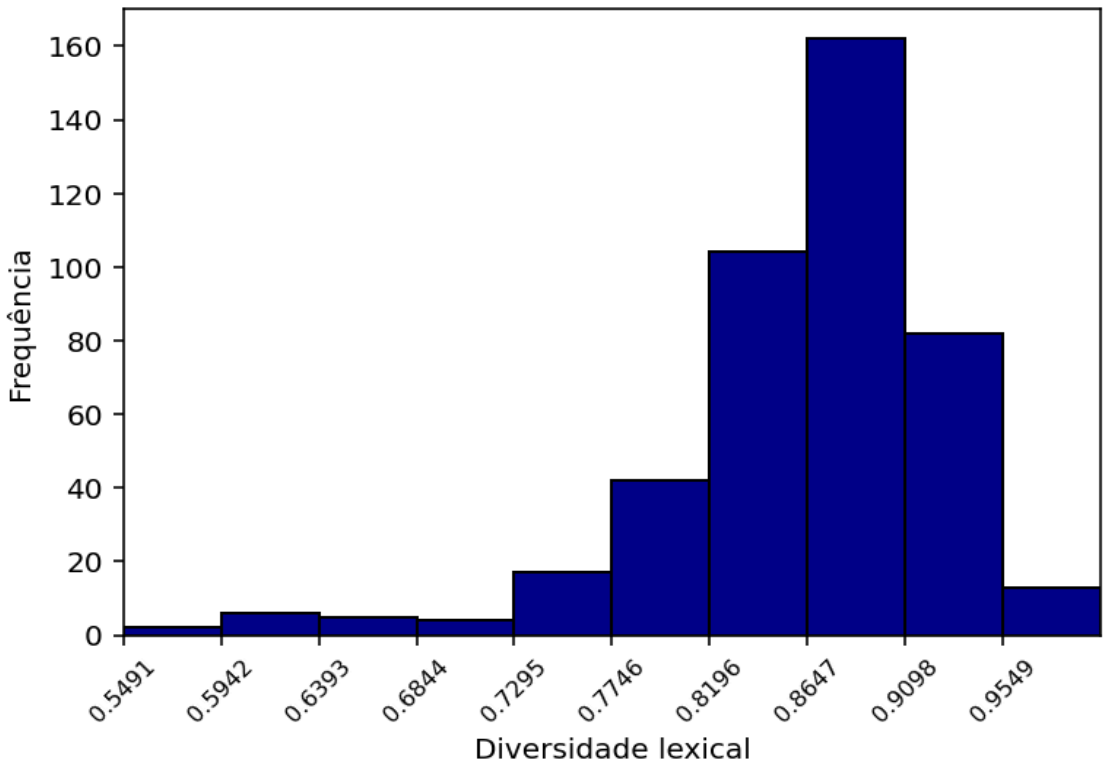

Figura B.8 *Distribuição da diversidade lexical sem as stopwords adicionadas*

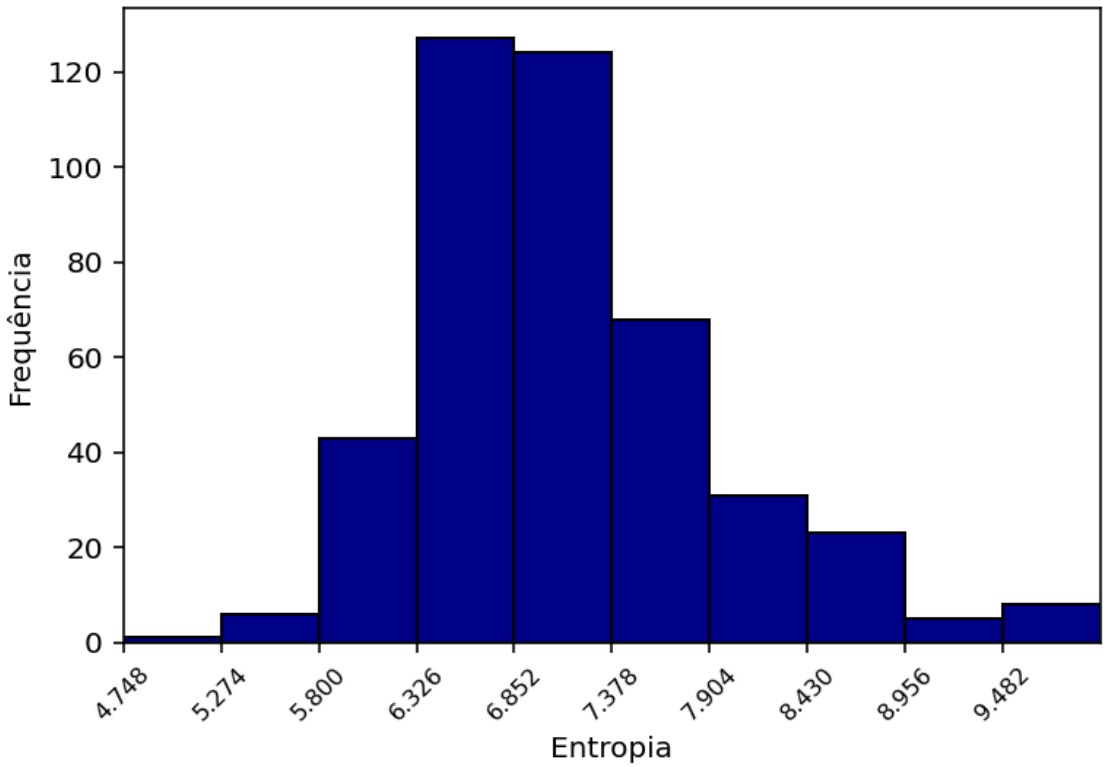

Figura B.9 *Distribuição da entropia com as stopwords adicionadas*

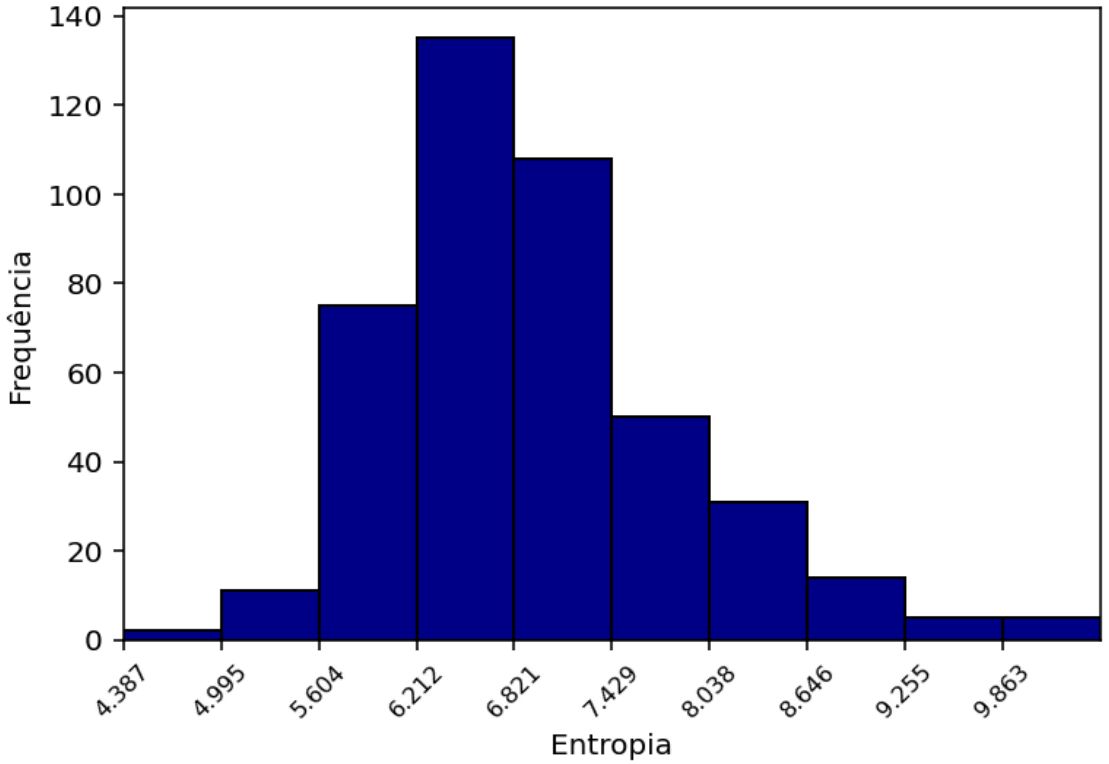

Figura B.10 *Distribuição da entropia sem as stopwords adicionadas*

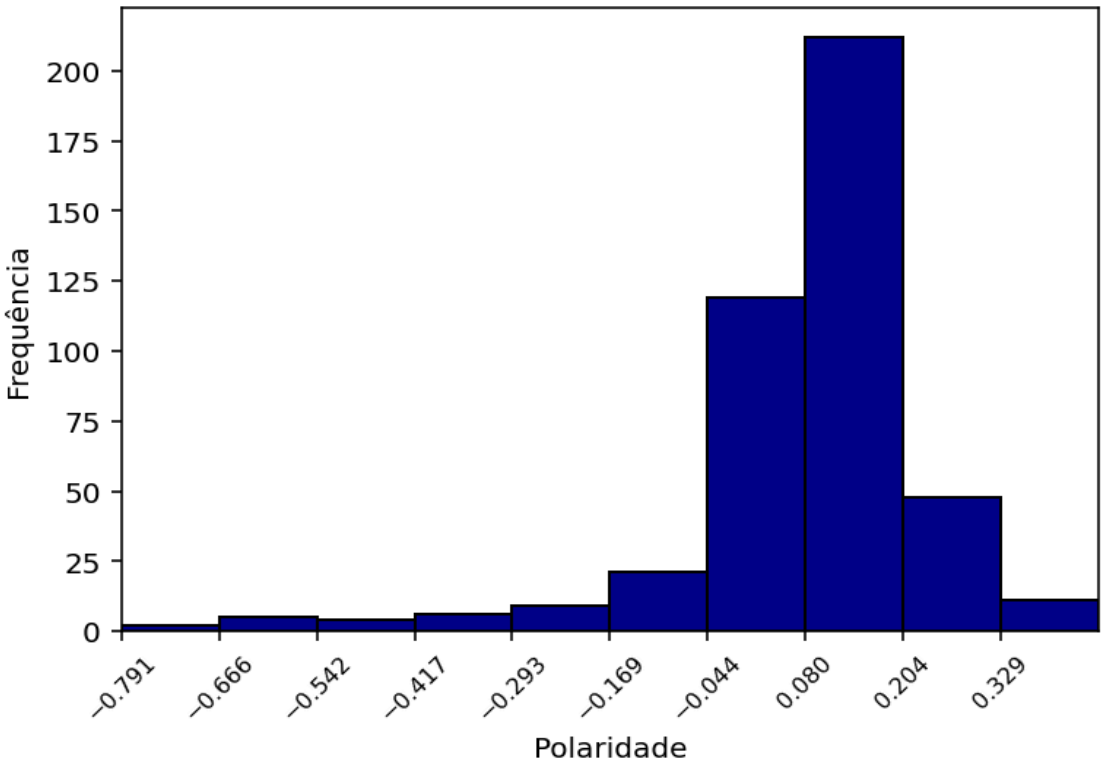

Figura B.11 *Distribuição da polaridade com as stopwords adicionadas*

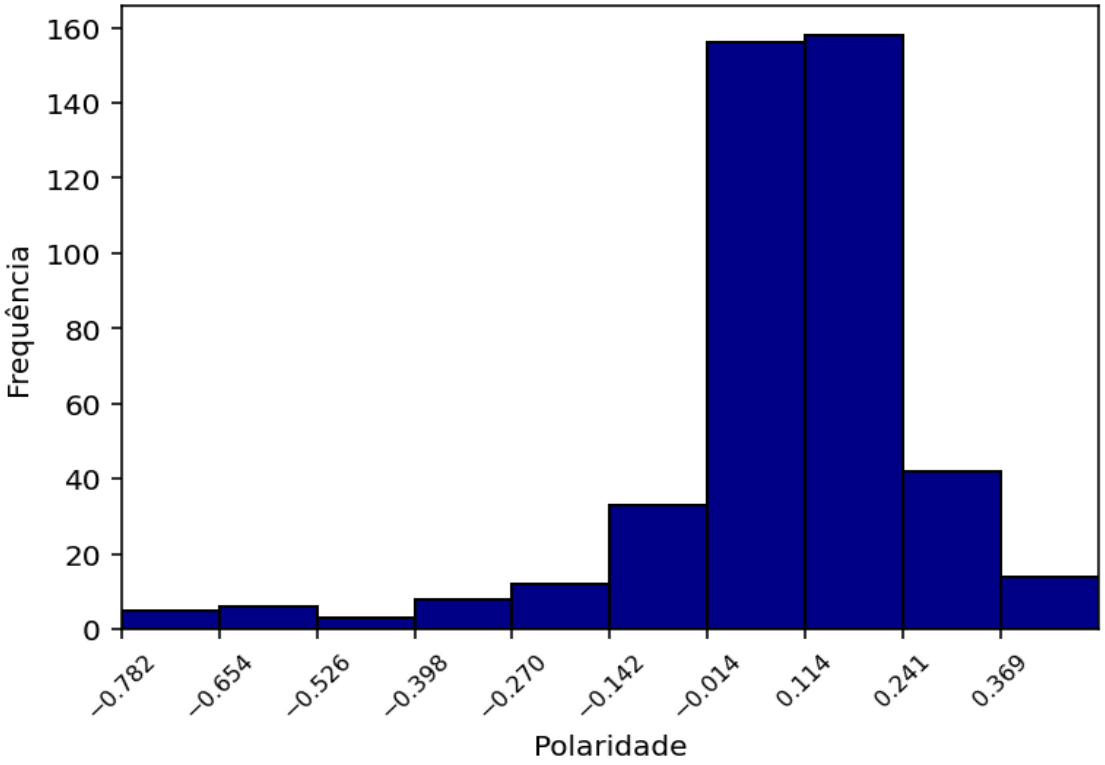

Figura B.12 *Distribuição da polaridade sem as stopwords adicionadas*

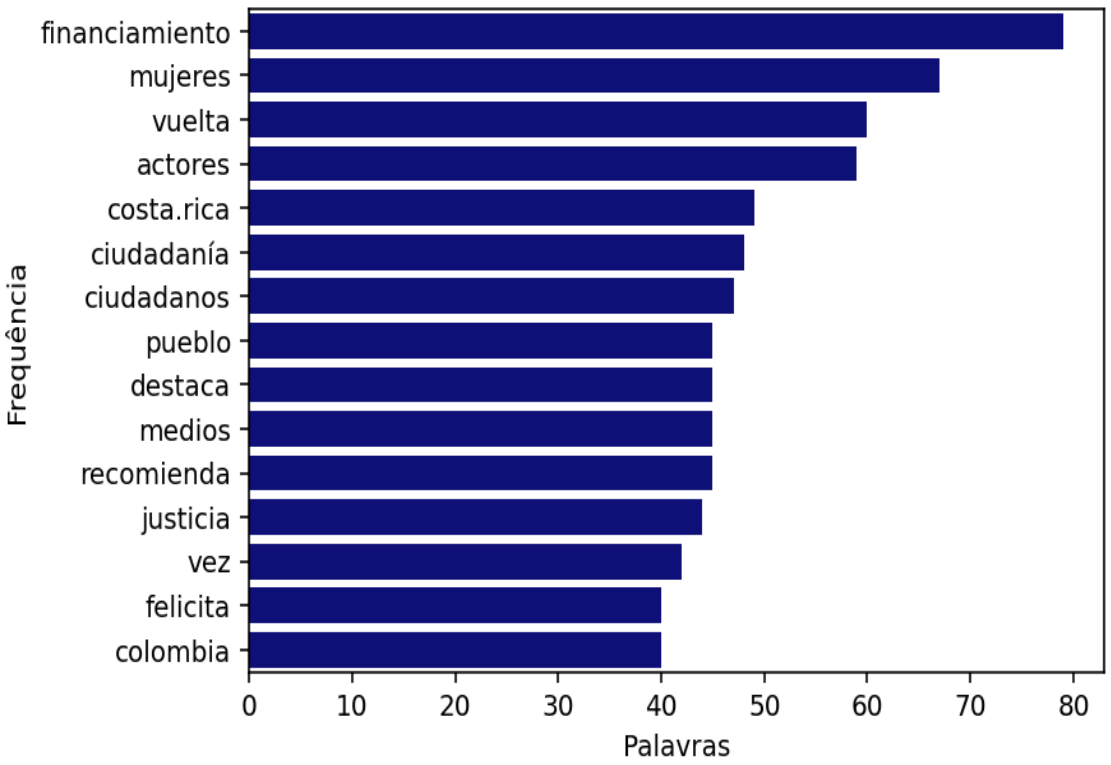

Figura B.13 *Palavras mais frequentes nos informes positivos*

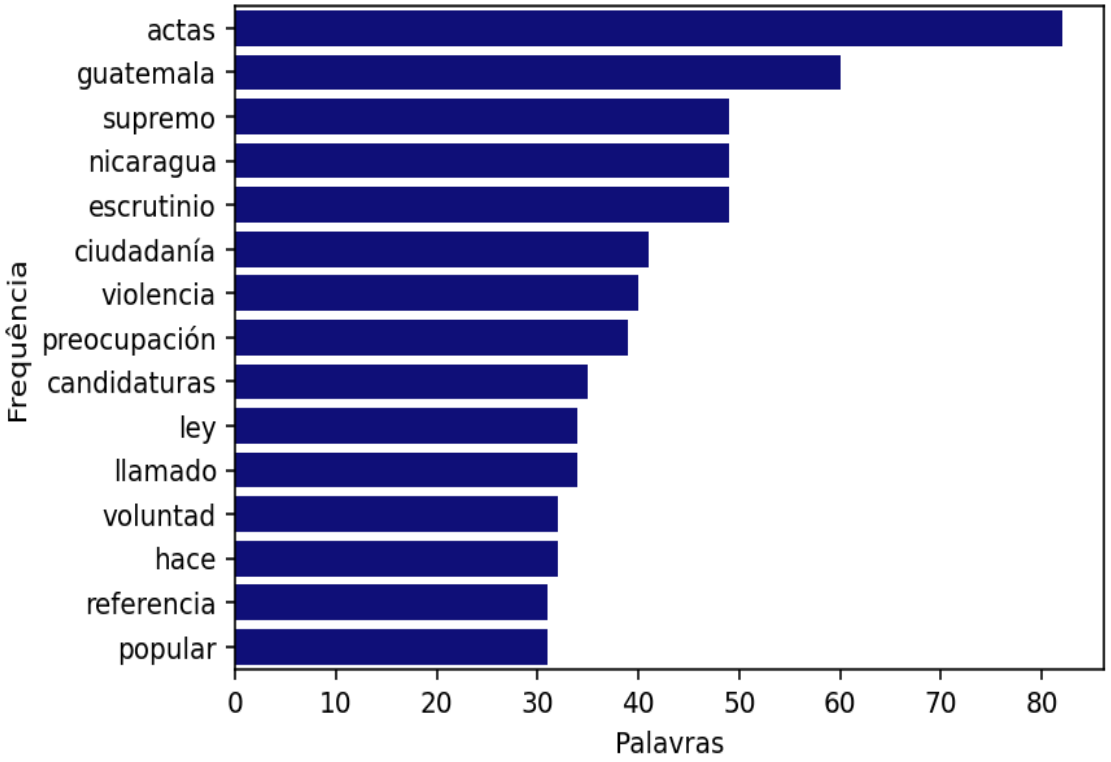

Figura B.14 *Palavras mais frequentes nos informes negativos*

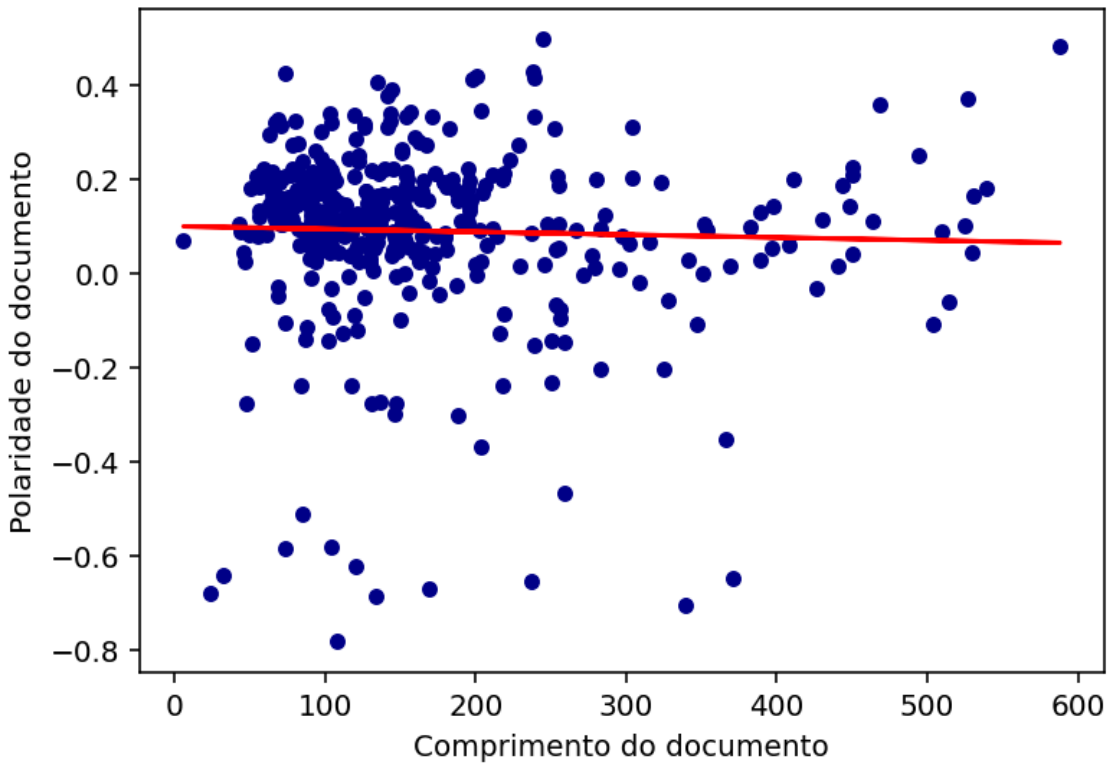

Figura B.15 *Correlação entre comprimento do documento e polaridade (ρ=-0,034)*

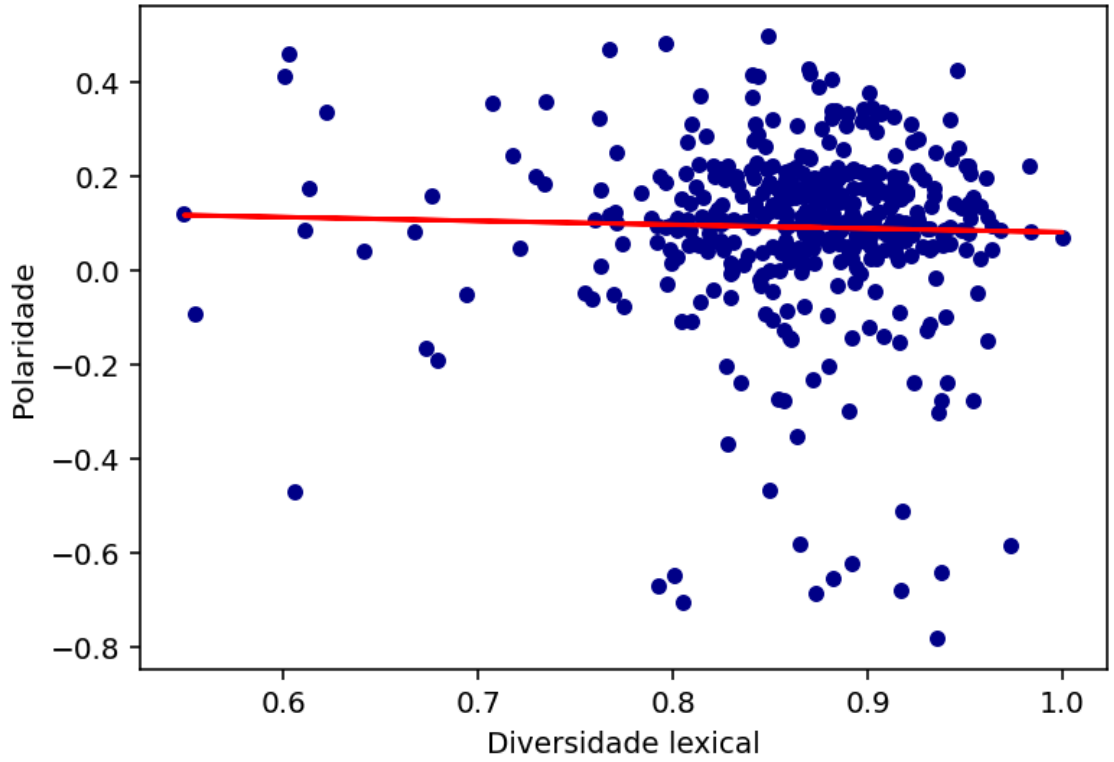

Figura B.16 *Correlação entre Diversidade lexical e polaridade (ρ=-0,0281)*

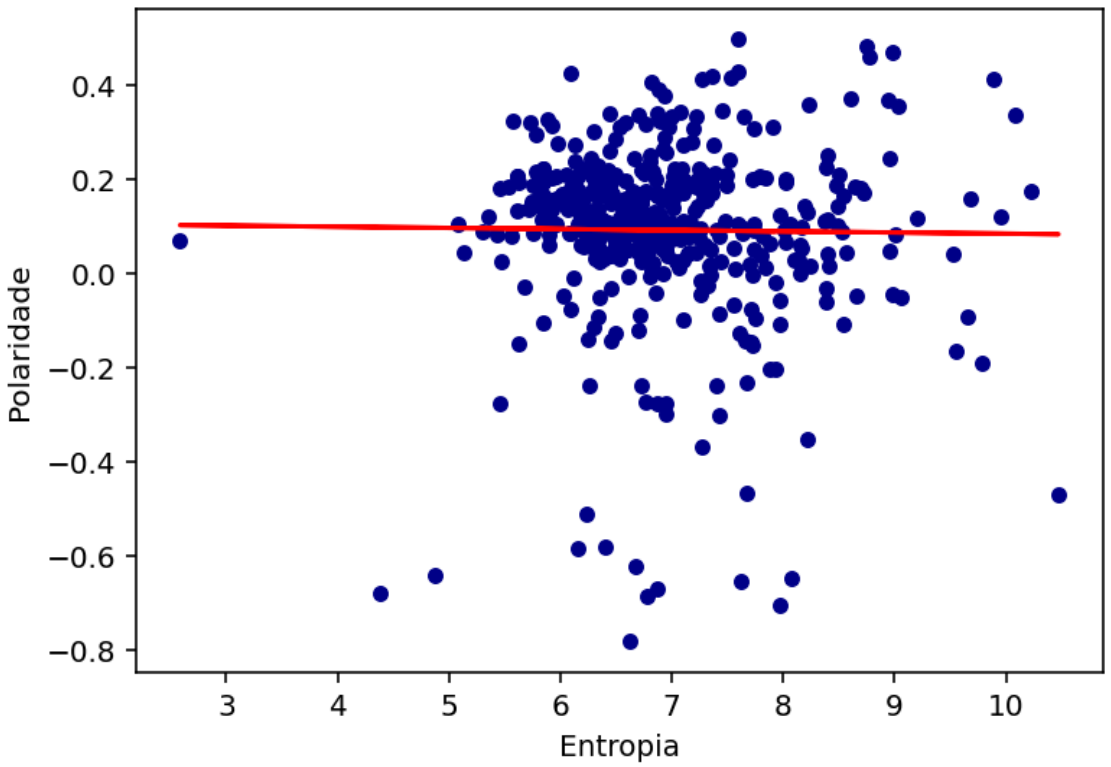

Figura B.17 *Correlação entre Entropia e polaridade (p=-0,0125)*

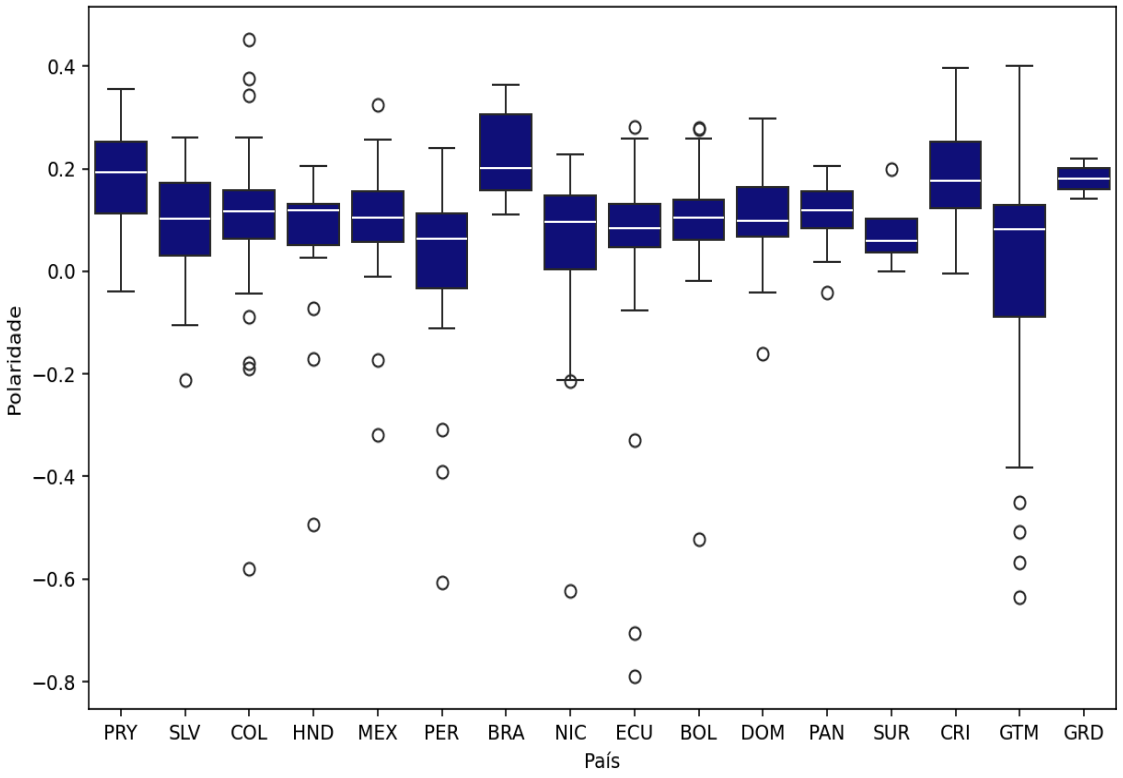

Figura B.18 *Box plot da polaridade dos informes por país de origem*

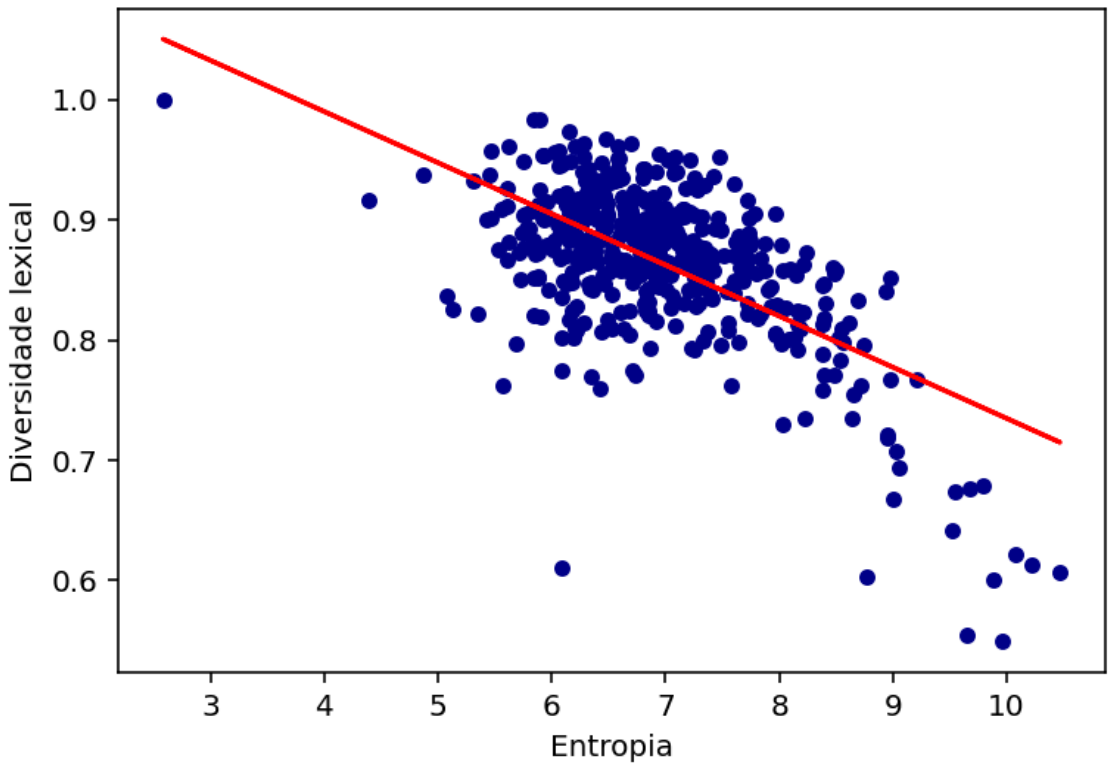

Figura B.19 *Correlação entre Entropia e Diversidade lexical (ρ=-0,6160)*

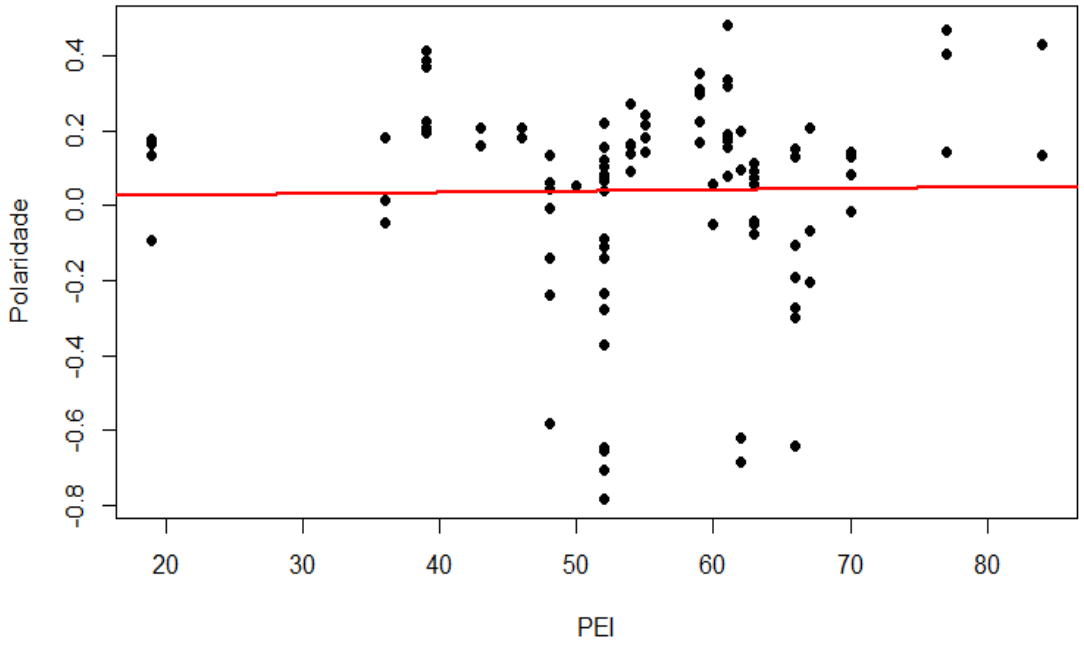

Figura B.20 *Correlação entre PEI e polaridade (ρ=0,0170)*

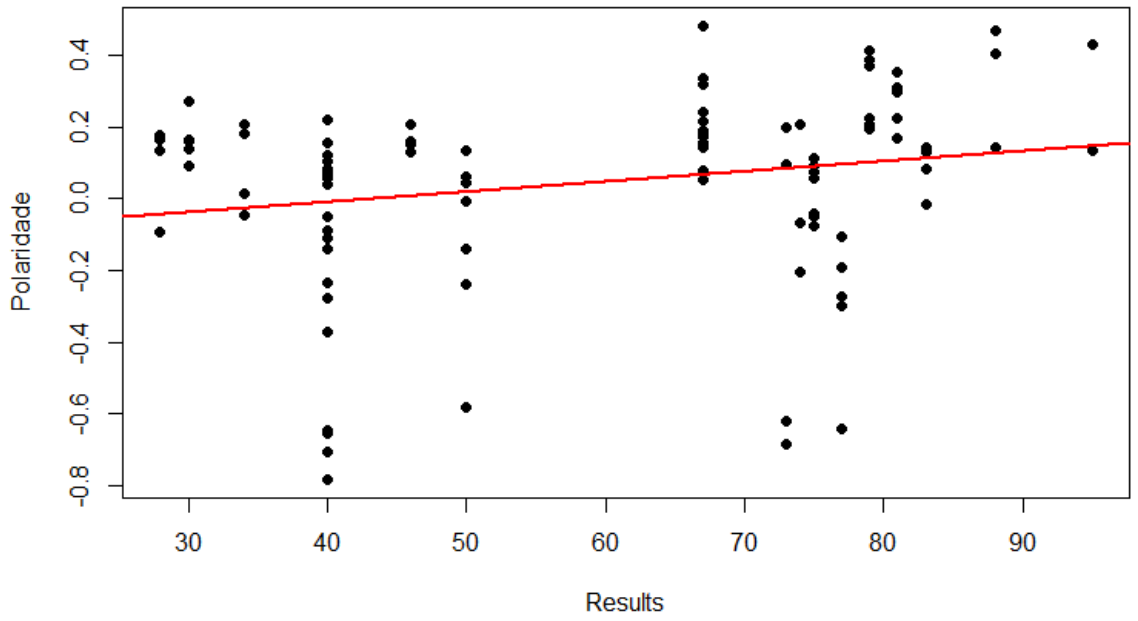

Figura B.21 *Correlação entre Results e polaridade (ρ=0,212)*

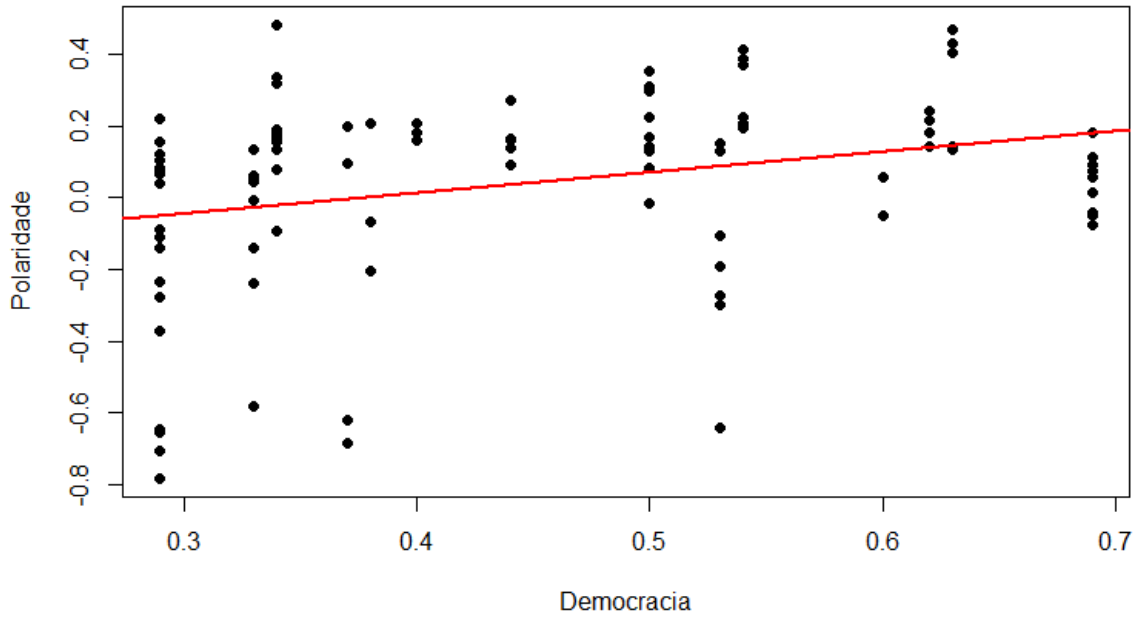

Figura B.22 *Correlação entre proporção de respostas "Democracia" e polaridade (ρ=0,2881)*

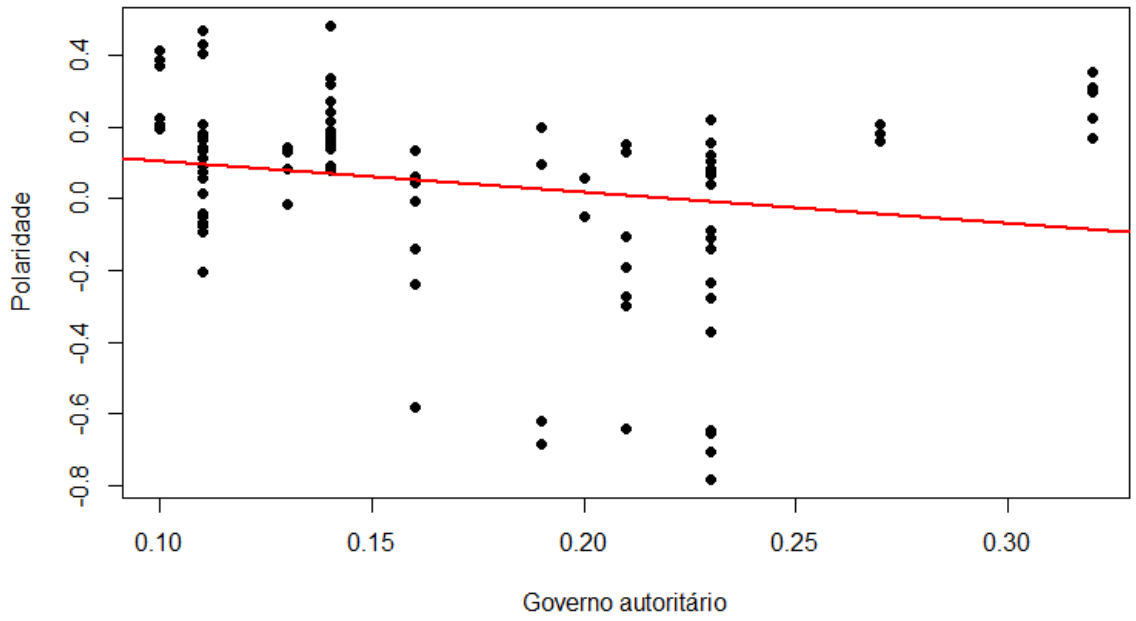

Figura B.23 *Correlação entre proporção de respostas "Governo autoritário" e polaridade (ρ=-0,201)*

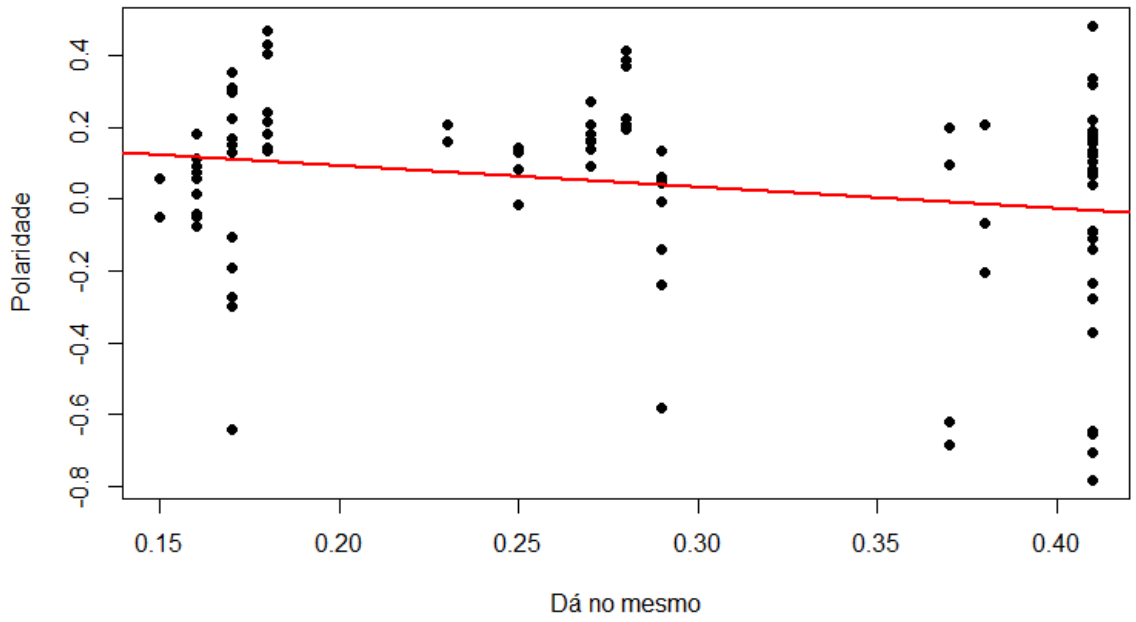

Figura B.24 *Correlação entre proporção de respostas "Dá no mesmo" e polaridade (ρ=-0,2241)*

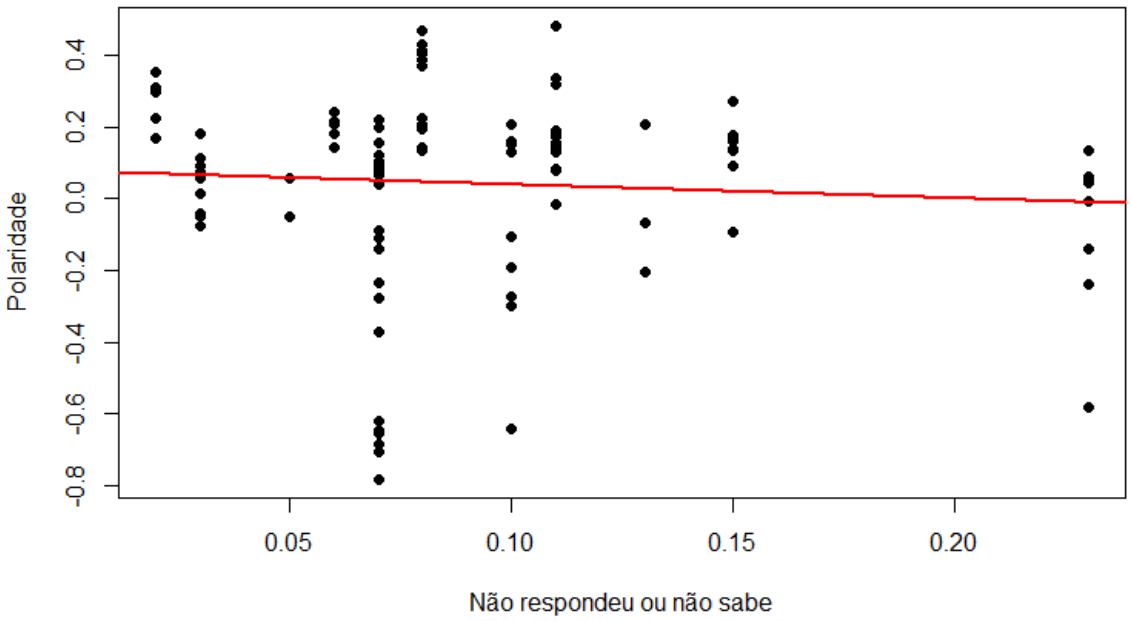

Figura B.25 *Correlação entre proporção de respostas "Não sei" ou não-respostas e polaridade (ρ=-0,0736)*

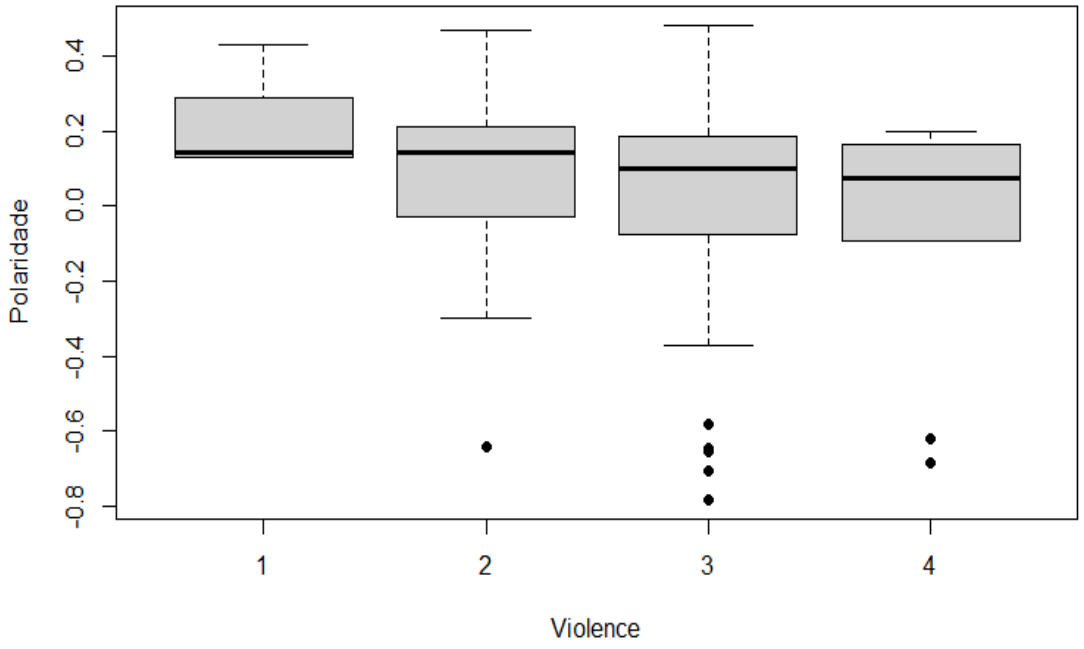

Figura B.26 *Box plot da polaridade por nível de Violence*

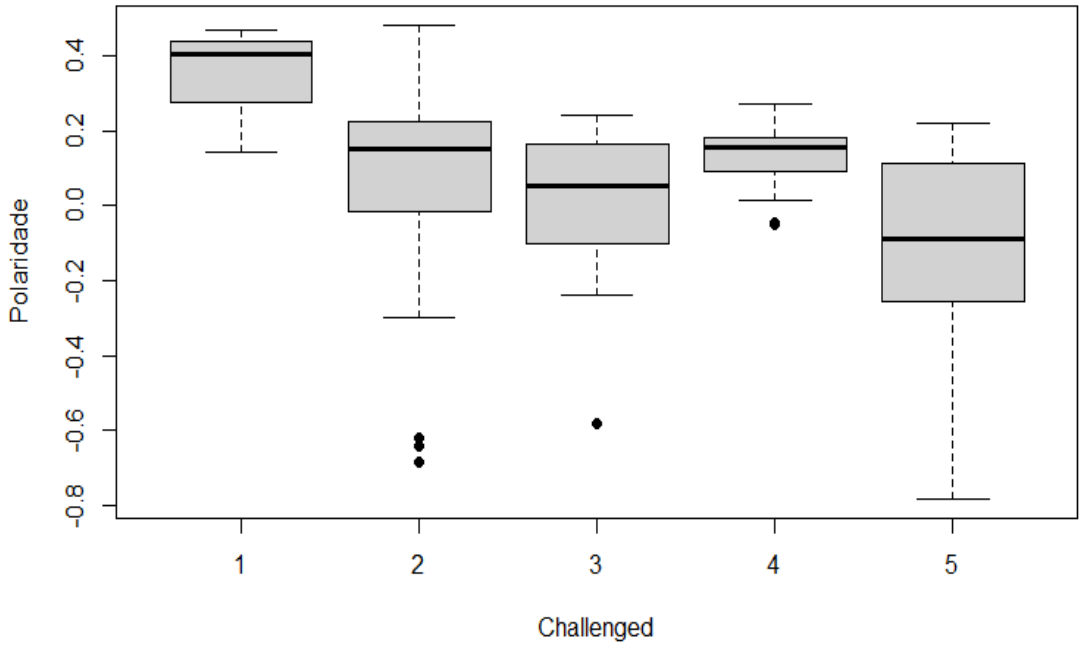

Figura B.27 *Box plot da polaridade por nível de Challenged*

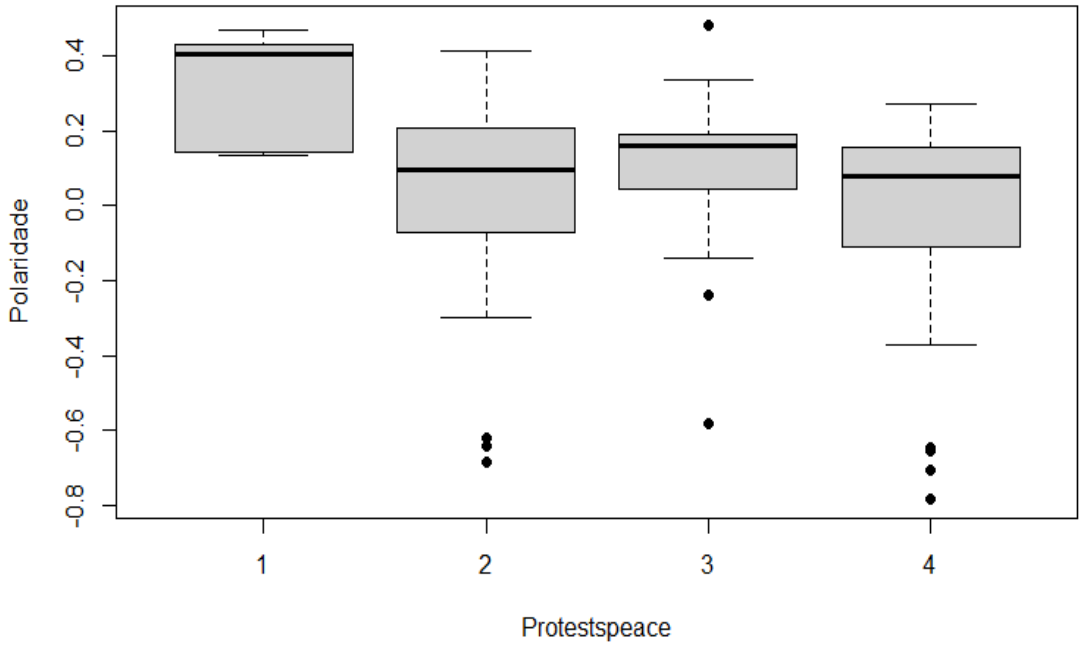

Figura B.28 *Box plot da polaridade por nível de Protestspeace*

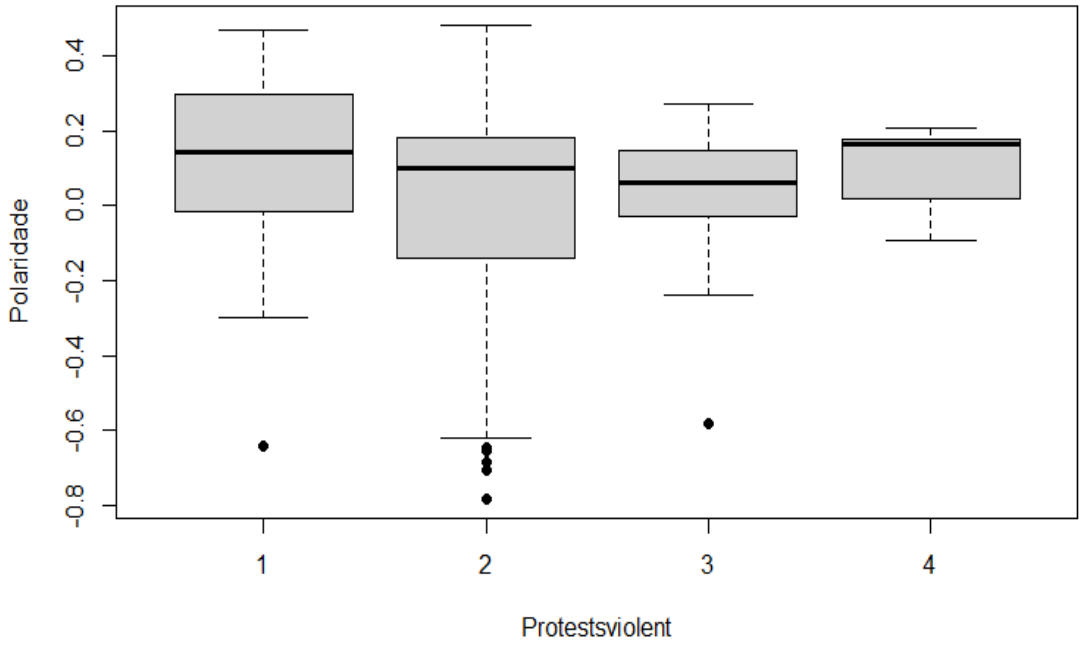

Figura B.29 *Box plot da polaridade por nível de Protestsviolent*

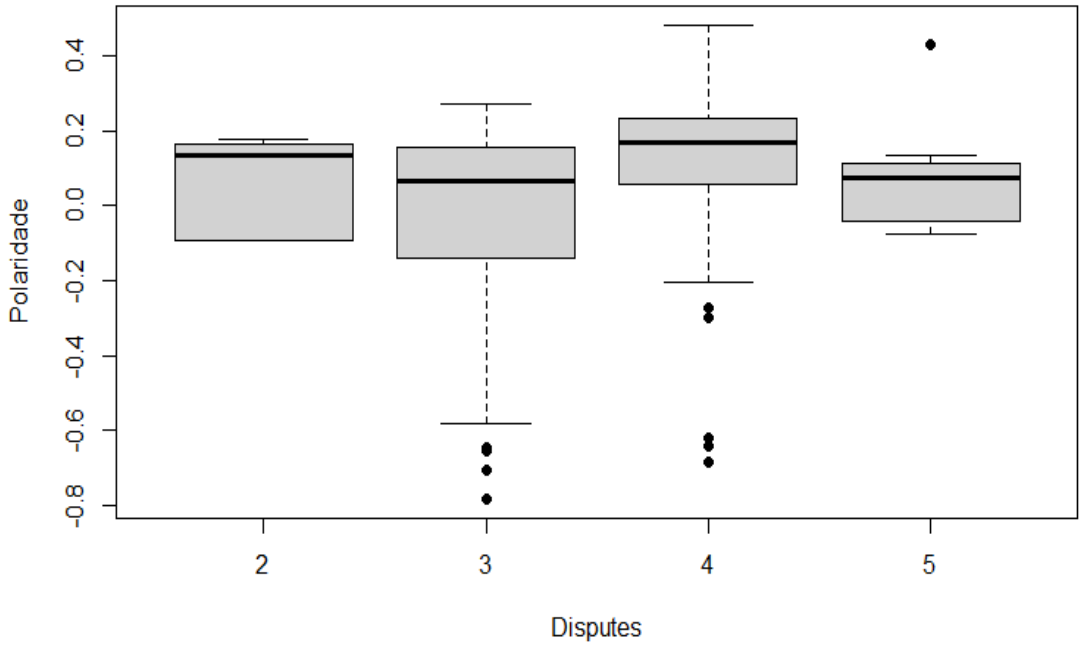

Figura B.30 *Box plot da polaridade por nível de Disputes*

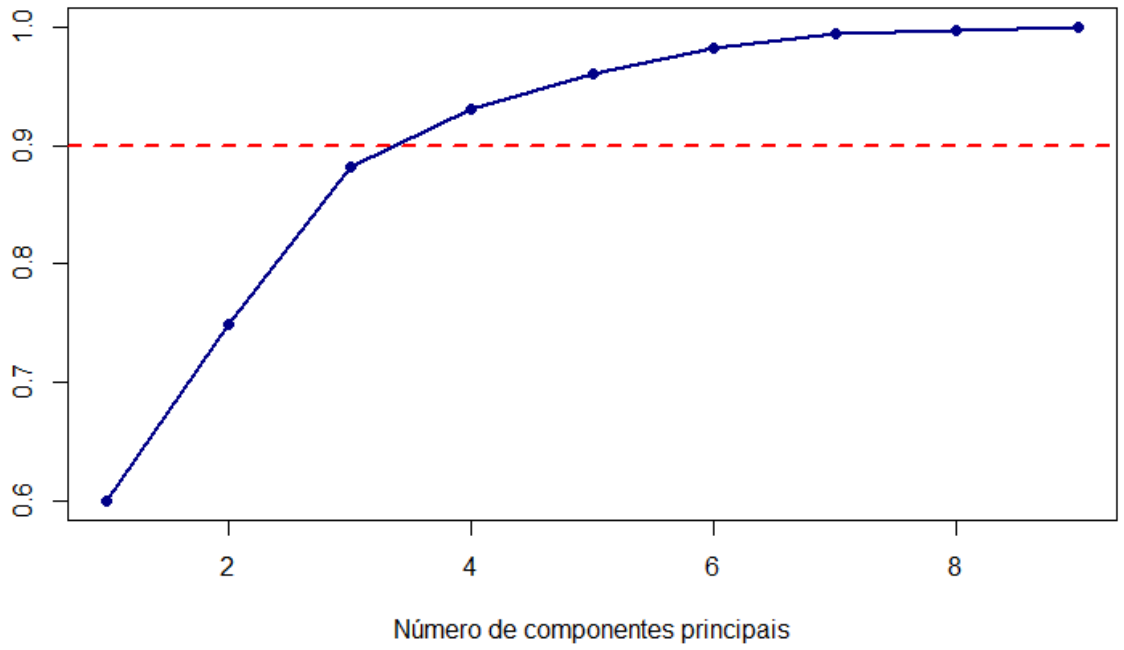

Figura B.31 *Variância explicada acumulada por número de componentes principais, com linha de corte em 0,9*

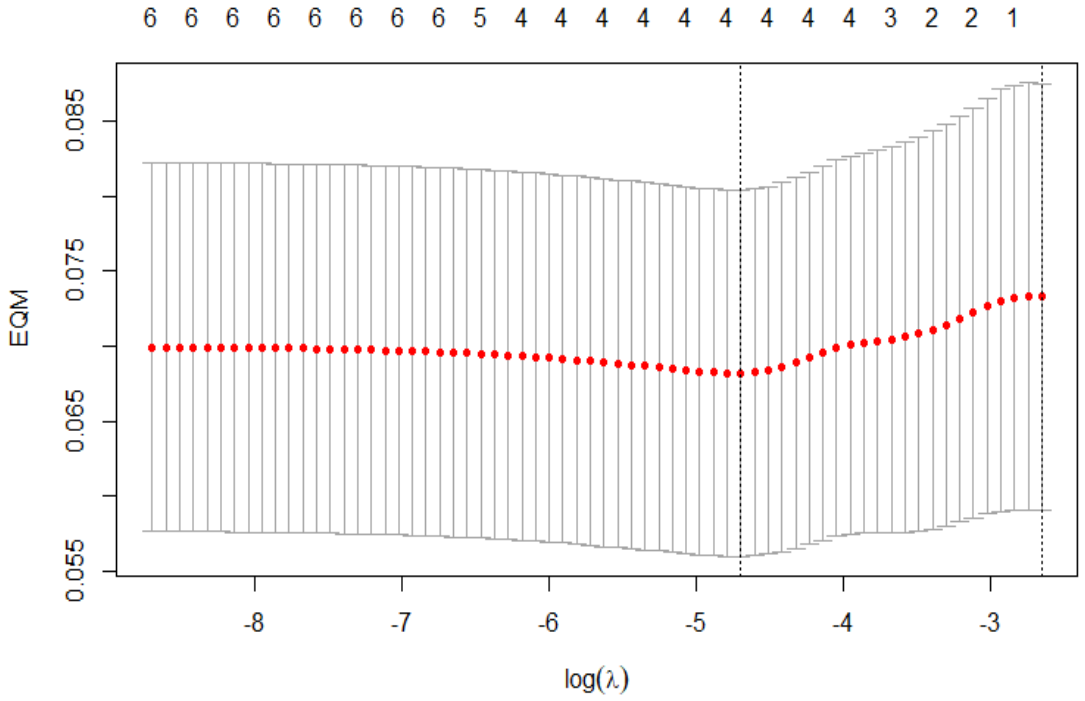

Figura B.32 *Erro quadrático médio da regressão LASSO por log(λ), com λ minimizador igual a 0,0091*