Investigating The Sentiment on Twitter After The Congress Invasion In Brazil

Giovanna Chaves, Lucas Costa Santos, Tobias "Tobinho" Pfeiffer

Barcelona School of Economics

March 27, 2023

Contents

1. Introduction
2. Methodology
3. Results
4. Conclusion & Future Research

Introduction

Context

- On January 8, 2023, a group of radical Bolsonaro supporters invaded the Brazilian Congress, Supreme Court and the Presidential Palace in Brasília.
- Brazil's political climate has been polarised since the elections (Oct 30, 2022) and leading up to them.
- In the West, Brazil politics has often been compared to the USA.

Figure: Bolsonaro Supporters Storming The Congress

Methodology

Outlook

Are people from Brasilia more affected by the insurrection than in other cities?

- Are they more politicised?
- Are they more positive/negative?
- Are they more critical of Bolsonaro?

Pipeline

1. **Scrape Twitter**
2. **Preprocess Text**
3. **Classify Tweets**
   3.a LDA
   3.b Naive Bayesian & Random Forest
4. **Regression Analysis**
   4.a Synthetic Difference-in-Difference
   4.b Difference-in-Differences

Scraping Twitter

- Dates: 02/01/2023 - 14/01/2023
- 5 cities:
  - Brasília
  - Sao Paulo
  - Rio de Janeiro
  - Fortaleza
  - Salvador

Challenges:

- Not enough control cities for Synthetic DiD
- 800K tweets to pre-process:
  - Computationally intensive
  - Too large to run through the LDA

Prepocessing Tweets

- Lowercasing
- Removing punctuation, diacritics and links
- Removing stopwords
- Tokenizing

Challenges:

- Translator functions are not fully developed yet
- Most resources (dictionaries, corpuses, etc.) are in English
- Keeping special characters
- Twitter is full of informal language and abbreviations

Classification with LDA

- LDA (Latent Drichlet Algorithm)
- Generative probabilistic model for text
- Discover hidden topics in a corpus
- Treats each document is a mixture of topics, and each topic is a mixture of words
- Uses MCMC

Classification with Naive Bayes and Random Forest

- Used dictionary to get political tweets
- Manually classified 747 tweets by their political bias
- Trained model on 522 of these tweets to test on the other 225
- Predicted on the rest of the dataset

Challenges:

- Dictionary might not adequately capture political tweets
- Enumerator biases
  - Classified independently and considered non-matching classes as neutral
- Very small proportion of classified tweets (too time-consuming)

Results

Tweet Classification (LDA)

- Sample 50,000 Tweets
- Choose 10 Topics in hyperparameters
- Initiate prior (one political & one negative)

How does LDA model topics?

- Approx. 4% of Tweets respectively were negative and politicised

Tweet Classification (Naive Bayes & Random Forest)

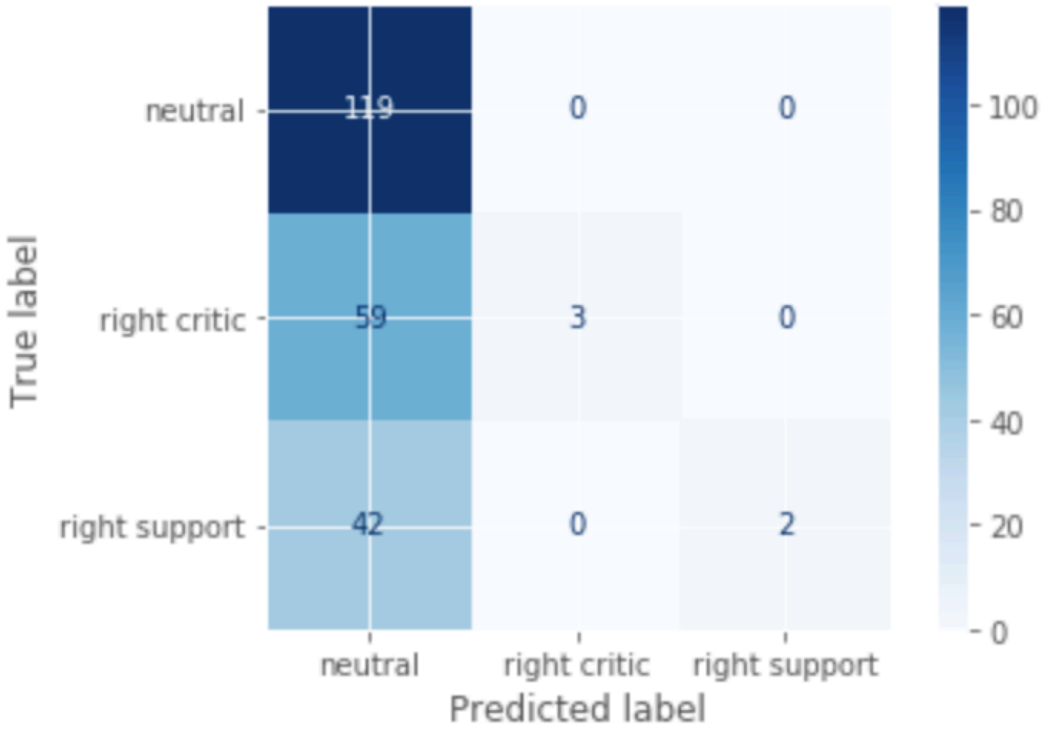
- High precision for anti-Bolsonaro and pro-Bolsonaro tweets in MNB, but very low recall...
- High recall for neutral tweets.
- More balanced with RF, but still not great...

		Multinomial Naive Bayes			Random Forest		
	Metric	**1**	**2**	**Joint**	**1**	**2**	**Joint**
Neutral	Precision	0.52	**0.77**	0.54	0.59	0.55	0.60
	Recall	**1.00**	0.27	**1.00**	0.95	0.52	0.97
	F1-Score	0.68	0.40	0.70	0.73	0.53	**0.74**
Anti	Precision	**1.00**	0.48	**1.00**	0.73	0.52	0.67
	Recall	0.12	**0.97**	0.05	0.34	0.79	0.26
	F1-Score	0.22	**0.64**	0.09	0.46	0.63	0.37
Pro	Precision	**1.00**	**1.00**	**1.00**	0.76	0.73	0.90
	Recall	0.06	0.17	0.05	**0.27**	0.24	0.20
	F1-Score	0.12	0.29	0.09	**0.39**	0.36	0.33

Figure: Performance metrics across classifiers.

Tweet Classification (Naive Bayes & Random Forest)

- Approx. 0.3% of tweets were anti-Bolsonaro and 0.06% pro-Bolsonaro with Naive Bayes.
- Approx. 1.8% of tweets were anti-Bolsonaro and 0.3% pro-Bolsonaro with Random Forest
- Strategy: only classify as anti or pro tweets those which were are certain



Regression Analysis (SDID/LDA)

- What is *Synthetic* Difference in difference?

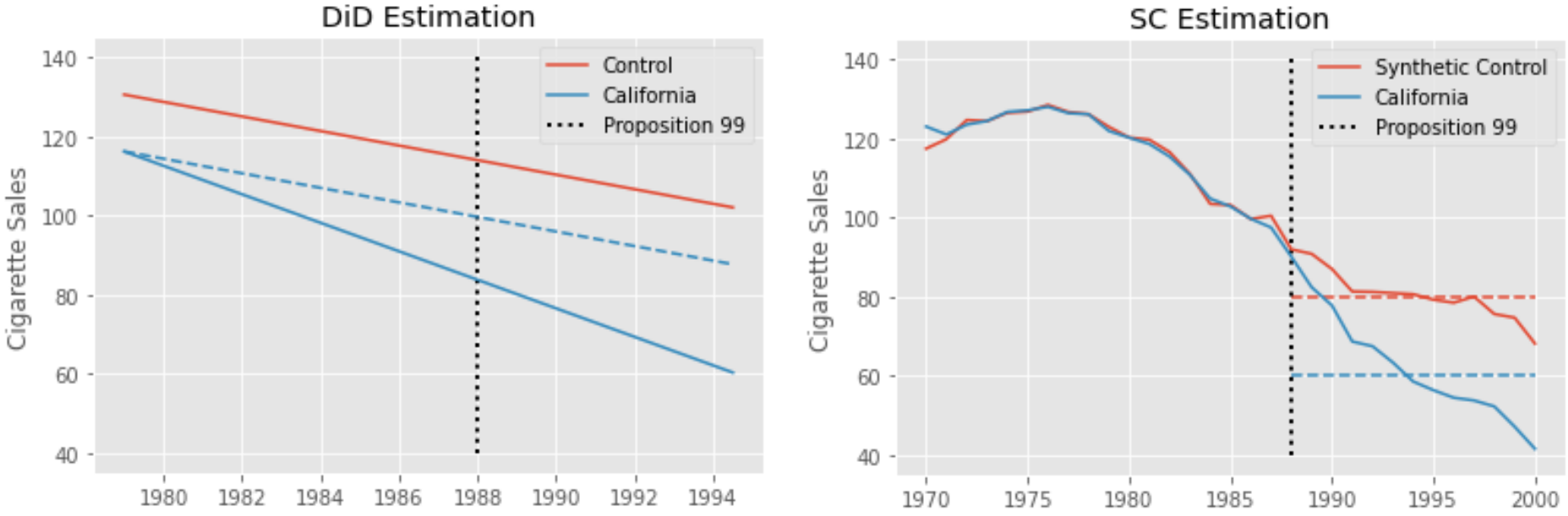


Figure: Intuition behind SDID

Regression Analysis (SDID/LDA)

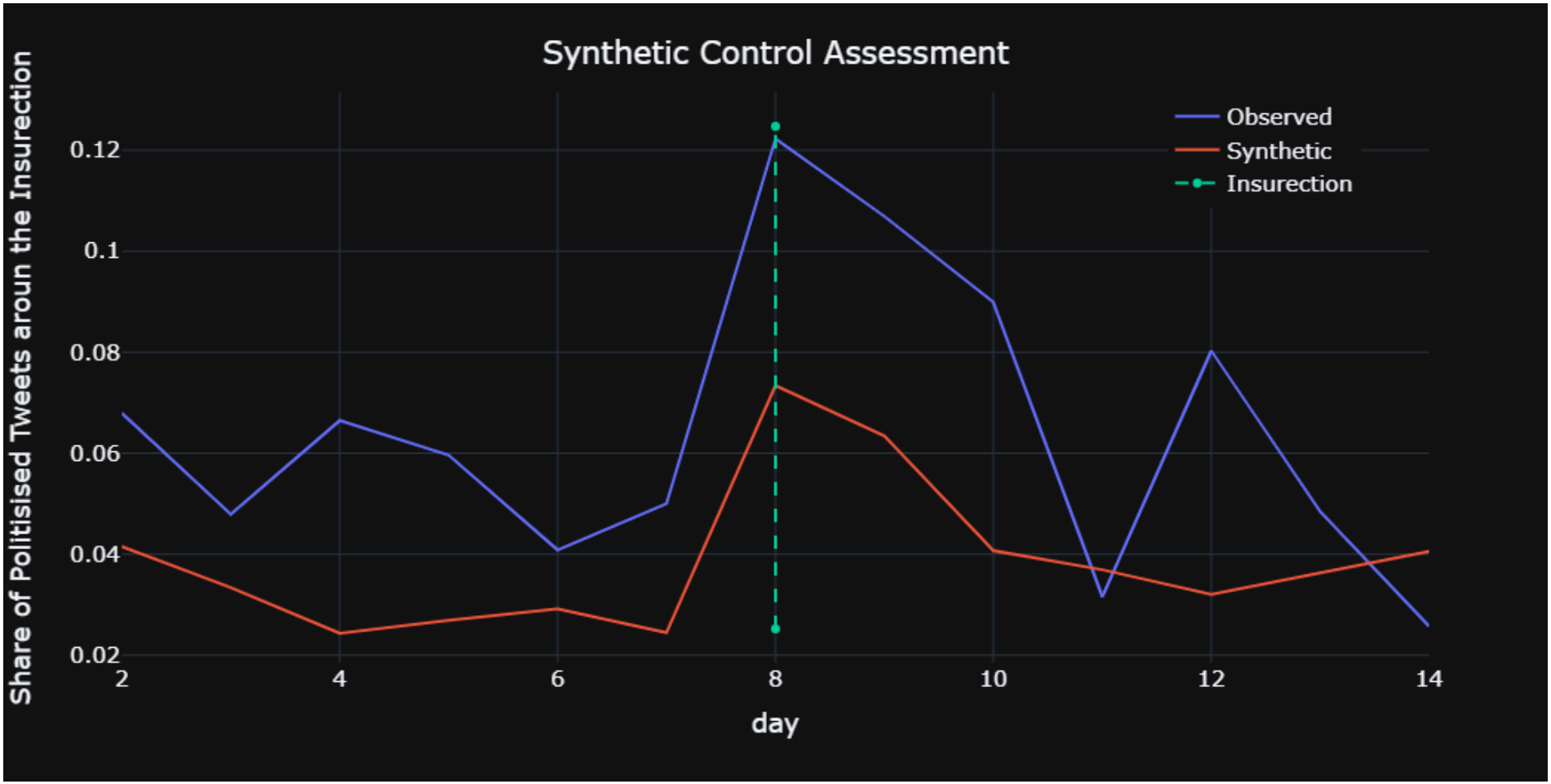


Figure: SDID for politicised Tweets, ATE: 0.08

Regression Analysis (DID/LDA)

$$P_{it} = \beta_0 + \beta_1 \cdot \text{Time}_{it} + \beta_2 \cdot \text{Treated}_{it} + \beta_3 \cdot \text{Time}_{it} \cdot \text{Treated}_{it} + \epsilon_i + u_t + v_{it}$$

OLS Regression Results

Dep. Variable:	political_share	**R-squared:**	0.333
Model:	OLS	**Adj. R-squared:**	0.307
Method:	Least Squares	**F-statistic:**	12.98
Date:	Tue, 21 Feb 2023	**Prob (F-statistic):**	2.68e-05
Time:	19:11:47	**Log-Likelihood:**	137.46
No. Observations:	55	**AIC:**	-268.9
Df Residuals:	52	**BIC:**	-262.9
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.0388	0.003	12.577	0.000	0.033	0.045
treated_time[T.True]	0.0154	0.006	2.482	0.016	0.003	0.028
treated[T.True]	0.0167	0.009	1.876	0.066	-0.001	0.035
treated_time[T.True]:treated[T.True]	0.0154	0.006	2.482	0.016	0.003	0.028

Omnibus:	2.288	**Durbin-Watson:**	1.335
Prob(Omnibus):	0.319	**Jarque-Bera (JB):**	1.423
Skew:	0.313	**Prob(JB):**	0.491
Kurtosis:	3.479	**Cond. No.**	4.47e+15

Figure: ATT $= 0.015$ (significant)

Regression Analysis (DID/LDA)

- Effect of the insurrection on the share of negative tweets violates the parallel trends assumption.

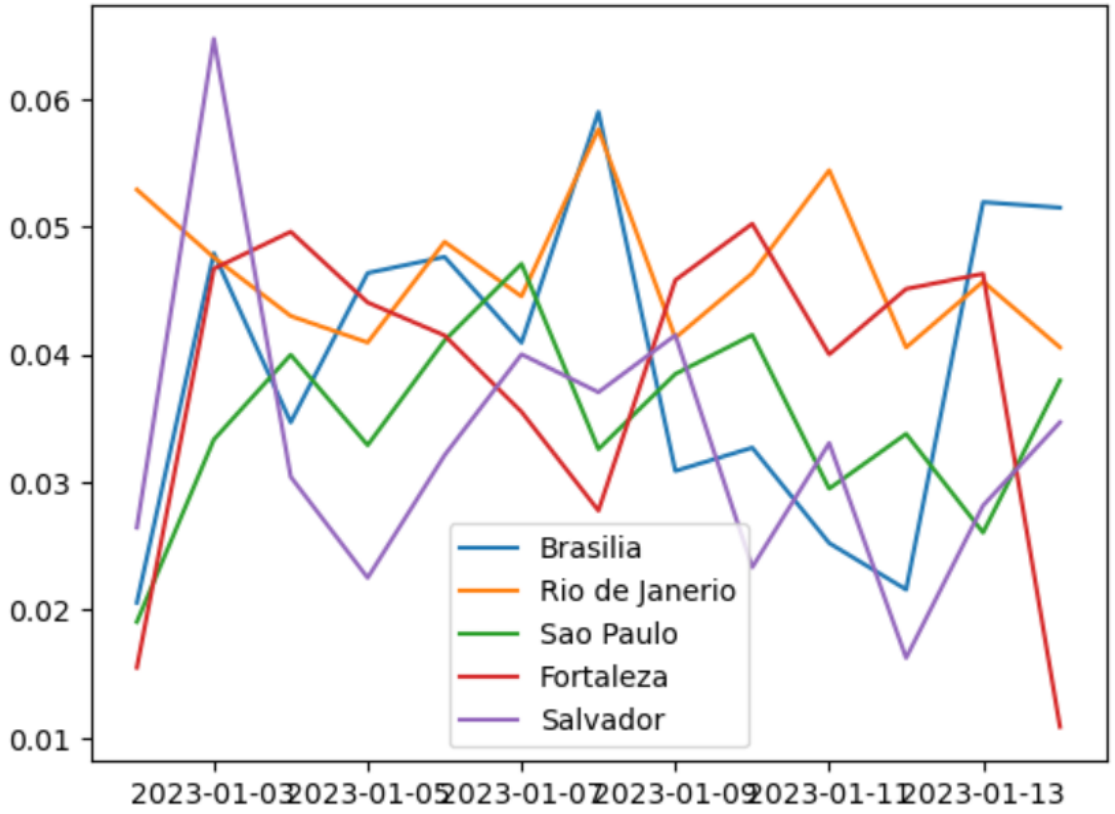


Figure: Share of negative tweets over time

Regression Analysis (DID/MNB)

- Effect of the insurrection on the share of anti-Bolsonaro tweets violates the parallel trends assumption.

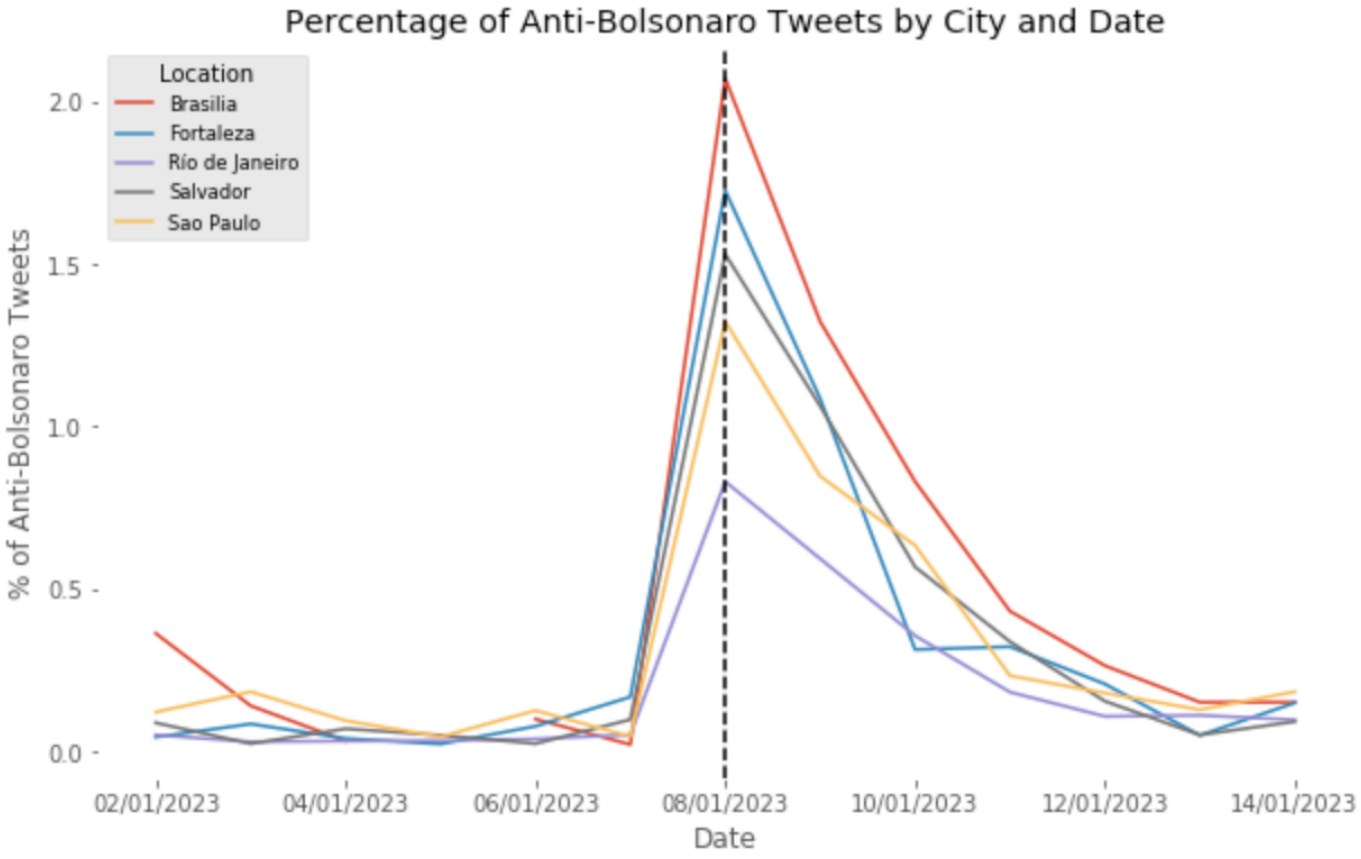


Figure: Share of Anti-Bolsonaro tweets over time

Conclusion & Future Research

Conclusion

- Challenges in adequately pre-processing such a large number of tweets not in English
- Used LDA to identify political and negative tweets
- Trained a classifier to predict political bias
  - Precision vs. recall trade-off
- No effect on negative share or anti-Bolsonaro bias (lower bound)
- 1.5% increase in political tweets in Brasília as a result of the insurrection.

Obrigado Pela Sua Atenção