Final Project

Giovanna Chaves, Lucas Santos and Tobias Pfeiffer

March 27, 2023

1 Introduction

On January 8, 2023, a group of radical Bolsonaro supporters invaded the Brazilian Congress, Supreme Court and the Presidential Palace in Brasília. In many ways, this invasion mirrored the January 6, 2021 attack on the United States Capitol, both motivated by increasing political polarization and election results that did not favor their preferred candidates. In this paper, we investigate the effects of this invasion on the political sentiment in the capital city, Brasília, by analyzing tweets from January 2023. We hypothesize that, compared to other major Brazilian cities, Brasília's tweets became more political, negative and critical of Bolsonaro and the far-right.

Our pipeline works the following way: first, we pre-process the text by lower casing, removing punctuation, removing stop words, and tokenizing tweets from the week before and the week after the Brasília invasion. Then, we use Latent Dirichlet Allocation (LDA) to classify tweets into specific topics, in order to identify those related to politics. Afterwards, we run LDA again to classify such tweets by sentiment: positive and negative. In parallel to the methods based on lexical resources, we use an alternative machine learning approach to identify political bias in tweets. First, we filter political tweets by using a dictionary of politically charged words, manually labelling a subset of these political tweets as supporting, opposing or being neutral towards Bolsonaro's government. The labelled data is then used to train a Support Vector Machine classifier to predict the political bias of the remainder of the data. Finally, we run a difference-in-differences estimation on the share of political tweets and probability of tweets critical of Bolsonaro in Brasília after the invasion, using the other four major cities as controls.

1.1 Literature Review

It is not unreasonable to assume that people from Brasília are more affected by the crisis that took place. While some might have directly taken part in the insurrection, others may have witnessed the attack and its aftermath in person. A study by the PewResearch Centre finds that Americans living in New York City and Washington D.C. report a higher personal impact than the nation average following 9/11 (Harting & Dorothey, 2021). This analogy supports our research design and ensure the relevance of the research question proposed in this work.

In the latest studies, to capture the sentiment and measure the impact of these events in general population, the social media, especially Twitter, is an common alternative. Pak and Paroubek (2010) were one of the first to study the validity of Twitter for a sentiment analysis: they argue that using the social media allows for a larger user base and picks up sentiments in real-time. The paper then presents a framework for sentiment analysis using Twitter, which includes techniques for pre-processing data, feature extraction, and sentiment classification, similar to the ones we use.

Tumasjan et al. (2010) also arrives in the conclusion that Twitter can clearly be considered as a valid indicator of the state of political opinion, and that it can complement traditional methods of conducting opinion polls. In a application of the advances area, the work of Ramasubbareddy et al. (2020) predicted the outcome of the 2016 US Presidential election with a high accuracy using a Twitter sentiment analysis.

Knowing that the social media can be used as a valid metric for sentiment analysis, in recent years the study of polarity in social media has become an important research topic in the field of natural language processing. Twitter, has consolidated among the academics as a popular platform for conducting such studies due to its vast user base and the wide range of topics discussed on the platform. One of the earliest studies on the classification of polarity in tweets was conducted by Go et al. (2009) who employed a supervised classification approach on tweets in English. However, constructing and manually tagging a corpus for the supervised classification of polarity is a difficult and expensive task, given the sheer volume of information published on Twitter. In order to overcome this challenge, the authors leveraged emoticons that usually appear in tweets to differentiate between positive and negative tweets.

The validity of this approach was demonstrated by Read, who showed that emoticons are reliable indicators of sentiment in tweets. The algorithms analyzed in this study were the same as those used by Pang et al. (2002), namely, Support Vector Machine (SVM), Naïve Bayes, and maximum entropy. These algorithms are commonly used in natural language processing tasks, and have been shown to be effective in the classification of polarity in Twitter data. In our study we compared the performance of the first two and a Random Forest model. The Random Forest model with modern word-embedding techniques like Word2Vec compared to traditional methods like TF-IDF, improve the accuracy of sentiment analysis, as can be seen in Hitesh et al. (2019).

In summary, the study of polarity in Twitter has gained significant attention in recent years, with many studies focusing on developing effective methods for classifying sentiment in tweets. The approach employed by Go, Bhayani, and Huang using emoticons as a proxy for sentiment classification has been shown to be effective and has been widely adopted in subsequent studies. However, due to the nature of tweets, containing irony, subjective content and informal language in majority of the time, the classification methods faces challenges, especially those studies conducted in the not most popular languages.

2 Data

We collected 888,416 tweets from Brasília, Salvador, Fortaleza, São Paulo e Rio de Janeiro between January 2nd, 2023 and January 15th, 2023. We chose these particular date as they fall one week before and one week after the invasion that happened in Brasília on January 8th. January 1st was excluded as that was the day in which Lula was sworn in as president in Brasília, which would dramatically increase the number of political tweets being posted across the country and, specifically, in the city on that day.

The other four cities were chosen as, together, they make up the five most populous cities in the country according to IBGE (Brazilian Institute of Geography and Statistics).

2.1 Data Collection

We used twarc2 to collect tweets from our desired sample, with the following command:

twarc2 search –archive "place:city_ID" –start-time 2023-01-02 –end-time 2023-01-15 city_1.jsonl

with city_ID referring to the place ID of each city. The jsonl files were then converted to csv for easier pre-processing.

2.2 Data Pre-Processing

Rather than using an all-encompassing function such as gensim's *simple_preprocess*(), we created our own pre-processing function. This was the easiest way to ensure that the pre-processing made sense for tweets in Portuguese, which include their own set of specific stop words, accents and abbreviations.

First, we transformed all text to lowercase, removed numbers and transformed all paragraph breaks into white spaces, removing double white spaces after. We also transformed "*&*", which is how "&" was being read, into "*e*". Then, instead of directly removing all punctuation, we removed usernames through regex, getting rid of @ and everything that comes after it until a white space. This ensures all usernames are removed, since directly removing punctuation would still leave usernames intact. This is also not a loss in information, since we keep columns *in_reply_to_username* and *quoted_username*, which already tells us if another user is mentioned in the tweet.

We then remove all other punctuation, as well as diacritics and links. Finally, we use NLTK's stop words in Portuguese and extend the dictionary with common slang and "chatspeak" that refers to the same stop words.

Initially, we attempted to lemmatize our tweets using spaCy's Portuguese pipeline. The benefit of lemmatizing on such a large dataset would be to improve the accuracy of the text classification and sentiment analysis by reducing the number of unique words to their base form. However, as most tweets are written in rather informal language, we found that spaCy's lemmatizer cannot capture the true meaning of the words. The lemmatizer only worked well with verbs, but it also required the verbs to be correctly written, which was not always the case. Other studies, including Thelwall, Buckley and Paltoglou (2011), used their own specially developed algorithm with the ability to correct spelling errors and understanding of the jargon often used on Twitter. It took too long to run the lemmatizer through our almost 1 million tweets. We found that the value added (in terms of reducing some verbs to base form) was very marginal compared to the computing cost, so we decided to just tokenize the words.

Similarly, a study by Go, Bhayani and Huang (2009) found that the use of POS-TAGS did not provide valuable information for the classification of polarity on Twitter, and that simply using unigrams still provides very good results. teste teste teste teste

2.3 Descriptive Statistics

From our sample of 888,416 tweets, the vast majority (approximately 52%) comes from Rio de Janeiro, followed by São Paulo (27%). While surprising, since São Paulo is the most populous city in the country, it could be explained by the fact that most of the population in the city travels to coastal cities for the summer holidays, with Rio being one of the most popular destinations.

Rio de Janeiro is also the city with the largest number of people tweeting daily, approximately

10 thousand unique authors per day, double that of São Paulo, the second largest. Once we consider the mean number of tweets per day and the mean number of unique tweet authors, we see that this is more or less balanced across the different cities: users in each city publish an average of between 3.44 (Rio) and 3.86 (Salvador) tweets per day (Table 1).

City	# Tweets	Mean Tweets/Day	# Unique Authors	Mean Authors/Day
Rio de Janeiro	462,595	35,584.23	31,573	10,353.23
São Paulo	242,360	18,643.08	17,901	5,037.23
Brasília	78,526	6,040.46	5,832	1,613.38
Salvador	50,801	3,907.77	3,724	1,012.00
Fortaleza	54,133	4,164.08	3,824	1,117.00

Table 1: Descriptive statistics for each city.

Finally, we plot the number of tweets (Figure 1) and users (Figure 2) in each city per day. Though there are slight fluctuations between cities, the general trend across time is quite similar: most cities see a slight increase in tweet volume between 02/01 and 05/01, with a drop until 07/01. This drop is much more pronounced in Rio de Janeiro and São Paulo than in Brasília, Salvador or Fortaleza. Interestingly, Brasília is the only city that observes a consistent fall in tweet volume during the first week of the year. This can potentially be explained by the number of people who travelled to Brasília for the presidential ceremony on the 1st and were likely still in the city on the 2nd.

All cities see a sharp increase in volume on the 8th of January, day of the invasion of Congress, Supreme Court and Presidential Palace. Rio de Janeiro pretty much recovers the tweet volume lost between the 6th and the 7th, as well as Salvador and Fortaleza, though for the latter this is much more subtle due to the fluctuation itself being smaller. São Paulo and, especially, Brasília, however, have a much higher gain in tweets on the day of the invasion. All cities see a small decline throughout the week after.

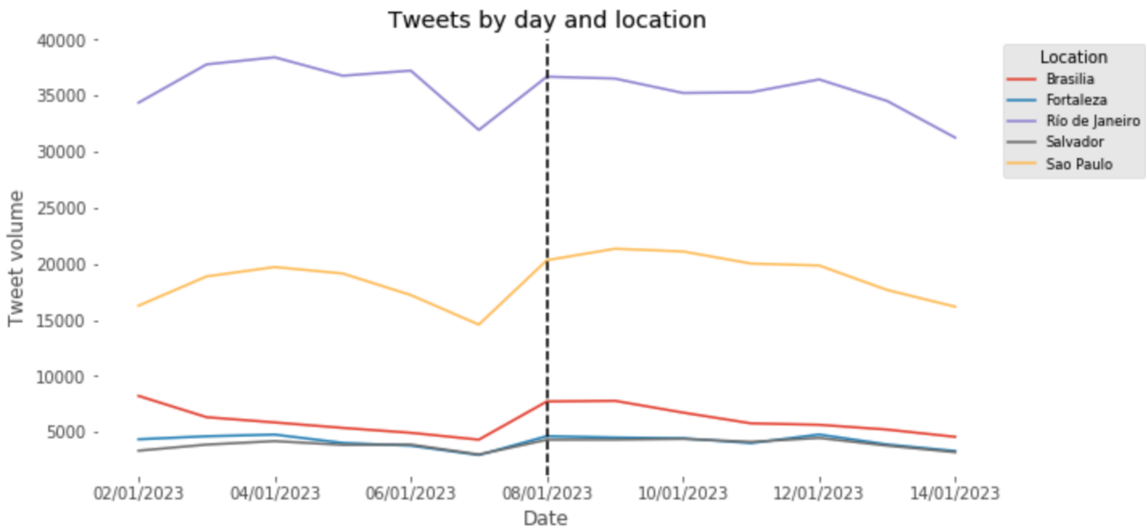

Figure 1: Volume of tweets posted daily in each city.

We also observe the same patterns in the number of unique authors tweeting across time, though that is much more constant than the number of tweets, especially for Fortaleza and Salvador.

From that, we can assume that there were few shifts in the number of people tweeting before and after the invasion, but a higher number of tweets per person on the day of the invasion, which would explain the sudden increase in cities like Brasília and São Paulo.

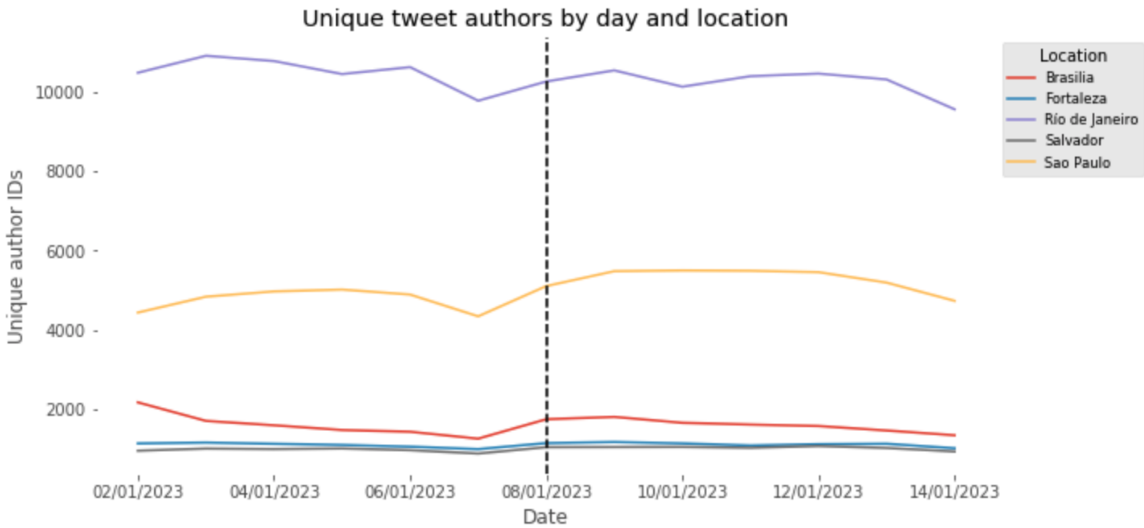

Figure 2: Number of unique authors tweeting daily in each city.

3 Methodology

We used a Latent Dirichlet Allocation to identify a political topic within the set of tweets, as well as to uncover clusters of positive and negative sentiment tweets within them.

Additionally, we experimented with training different classification models (Naive Bayes, Regression Forest, Support Vector Machines) to predict tweets' political bias.

3.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that is widely used in natural language processing. It is used to discover the hidden topics within a corpus of documents, making the assumption that each document within the corpus is a mixture of topics and each topic a mixture of words. LDA works by first randomly assigning each word in a document to a given topic, then adjusting the topics and words iteratively to maximize the probability of each document given the topics. At the end of the process, each document will have a set of topics assigned to it, and each topic will have a set of words associated with it. This allows for the discovery of underlying topics within the documents, which can then be used for various tasks such as document classification, topic modelling, and recommendation systems.

In this paper, we use LDA to identify tweets related to politics, as well as positive and negative-sentiment tweets, given the absence of an adequate opinion lexicon list for Portuguese.

3.2 Labelling the training data

In order to train both classifiers to analyze political bias, it was necessary to manually label a subset of tweets. For an unbiased data-labelling, two people independently classified 747 tweets into three categories: neutral tweets, pro-Bolsonaro tweets, and anti-Bolsonaro tweets.

These classifications were then combined to optimize performance. Where the two enumerators agreed, the label was kept. If the two enumerators disagreed, the tweet was deemed neutral. These labelled data were then used to train the SVC, Random Forest and Naive Bayes models, in order to predict the political bias of the remainder of the political tweets.

3.3 Classifiers

In the absence of a proper dictionary of positive and negative Portuguese words and subpar translation methods, we split our labeled tweets between train (522 tweets) and test (225 tweets) to train a classification model. We applied the Term Frequency-Inverse Document Frequency Vectorizer to build the pipeline for the models, and tuned hyperparameters with $GridSearchCV$.

We trained three different classifiers (Naive Bayes, Random Forest, and Support Vector Machine), though results are only presented for the model with highest ROC AUC score, which was the SVM.

We initially opted to use a Naive Bayes to predict the probability of tweets belonging to three different classes: neutral, pro-Bolsonaro and anti-Bolsonaro. This is an algorithm based on the Bayes theorem and with prior applications in the NLP field, including Pang, Lee and Vaithyanathan (2002) and Go, Bhayani and Huang (2009). It provides a simple, fast and good baseline, though its main weakness is its prediction accuracy, which was generally lower than other models.

Additionally, we compared its performance with a Random Forest classifier. The Random Forest model dealt well with high dimensional data (in our case the train dataset has 522 rows and 3234 columns) and reported a better performance than the Naive Bayes.

Finally, we ran a SVM classifier which turned out to be our best model in terms of performance. One of the main weakness is that could be computationally expensive train a model with a large dataset, but this wasn't a problem in our work since we had few training observation. The SVM performs well when there is a clear distinction between classes, which we believed to be the case for the political tweets.

3.4 Difference-in-Differences

We used the results of our LDA and SVM classifier to estimate the impact of the insurrection in Brasília on different outcomes: proportion of political tweets, proportion of negative/positive tweets, and likelihood of being a politically biased tweet. We do so through a difference-in-differences analysis, which is a quasi-experimental design that compares changes in outcomes over time between a group that receives the intervention and one that does not. In our case, we compare the tweets in Brasília to those in the other four major Brazilian cities, relying on the assumption of parallel trends, that is, that before the insurrection all five cities had similar trends in tweets.

Since we don't have covariance, we define the following models:

$$P_{it} = \beta_0 + \beta_1 \cdot \text{Treated}_i + \beta_2 \cdot \text{Time}_t + \beta_3 \cdot \text{Time}_t \cdot \text{Treated}_i + \epsilon_{it} \tag{1}$$

$$P_{it} = \beta_0 + \beta_1 \cdot \text{Treated}_i + \beta_2 \cdot \text{Time}_t + \beta_3 \cdot \text{Time}_t \cdot \text{Treated}_i + u_i + \epsilon_{it} \tag{2}$$

$$P_{it} = \beta_0 + \beta_1 \cdot \text{Treated}_i + \beta_2 \cdot \text{Time}_t + \beta_3 \cdot \text{Time}_t \cdot \text{Treated}_i + v_t + \epsilon_{it} \tag{3}$$

$$P_{it} = \beta_0 + \beta_1 \cdot \text{Treated}_i + \beta_2 \cdot \text{Time}_t + \beta_3 \cdot \text{Time}_t \cdot \text{Treated}_i + u_i + v_t + \epsilon_{it} \tag{4}$$

The first model is a basic DiD estimation, the second adds city fixed effects, the third adds time fixed effects, and the fourth includes both fixed effects.

4 Results

4.1 Detection of topics with LDA

We choose the LDA, since most existing classifiers only work well for the English language and may perform poorly on informal language. Note that the LDA only considers the frequency of words and therefore it tends to perform well in our settings.

Due to the long running time of the LDA-algorithm, we sampled 50,000 Tweets. For the politicised-classification we initiated the LDA-algorithm with 10 clusters. However, the first iteration did not yield a clear classification. That is, we obtained two clusters containing a significant share of words related to politics broadly and the insurrection specifically. To avoid, this we specified a prior that contained politically-charged words. Without a prior, all words in the corpus have uniform probability. By defining a dictionary with political words we up-weight the initial probabilities in comparison to the other words in the corpus. This technique allowed us to clearly distill a political topic. We then mapped the topic back to the sample of tweets, resulting in 11.5% (roughly 5,750) of the sample relating to politics.

We attempted a sentiment detection with the LDA by passing positive and negative words as our prior. However, the algorithm was only able to pick up one clear sentiment topic, which contained words related to negative feelings. Approximately 1,600 tweets in the sample were negative sentiment tweets, or roughly 3.2%.

4.2 Training the classifiers

We defined binary classifications for our models in order to more accurately measure the likelihood of being politically biased. That is, one model predicted the probability of being a tweet critical of Bolsonaro compared to not being a tweet critical of Bolsonaro (i.e. either supportive or neutral); the second predicted the probability of being supportive of Bolsonaro compared to not (i.e. either critical or neutral); the final model predicted the probability of being neutral compared to not (i.e. either critical or supportive).

Figure 3 shows the ROC AUC of the Support Vector Machine Classifier for the likelihood of being an opposition tweet, which turned out to be our best-performing model. This is our main variable of interest, since our hypothesis is that Brasília's tweets became more critical of Bolsonaro and the far-right after the insurrection, compared to other cities.

Once we had the model, we used it to predict on the remainder of our sample. This resulted in an average likelihood of 22.13% for being a tweet in opposition to Bolsonaro, 14.88% for being a tweet supporting Bolsonaro, and 63.91% for being neutral, across all cities during the period analyzed.

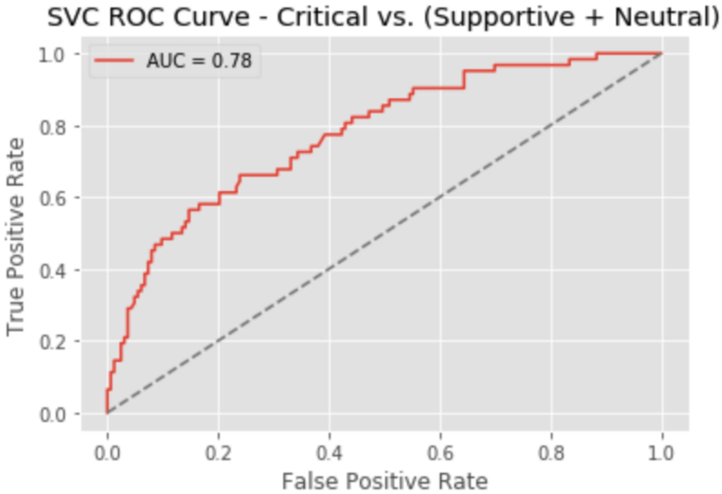

Figure 3: ROC AUC of the SVC model.

4.3 Difference-in-differences estimates

We estimated the effect of the insurrection on the share of political tweets and the probability that a tweet is critical of Bolsonaro. Although we initially hoped to estimate the effect on the share of negative tweets and likelihood of being supportive of Bolsonaro, these two outcomes violated the parallel trends assumption, one of the requirements to use a difference-in-differences estimator.

4.3.1 Effect on Share of Political Tweets

In this section we estimate the specifications of Section 3.4. Since the share converges to the same share, we cut off the last two days of the sample. All standard errors are robust against heteroskedasticity.

Our estimates indicate that, after the insurrection, the share of political tweets in Brasília increased by 1.1%, with a significance level of 10%. This effect is robust when we take into consideration time fixed effects, but is no longer significance in our final specification, which includes both city and time fixed effects. This could be due to the collinearity issues that arose once we added city fixed effects.

4.3.2 Effect on Likelihood of Political Bias

We also run our four models on the likelihood of the tweet being opposed to Bolsonaro (Table 3) or neutral towards Bolsonaro (Table 4).

Our estimates indicate that, after the insurrection, the likelihood of a tweet in Brasília being critical of Bolsonaro increased by 1.47 percentage points, with a significance level of 5% for our more robust estimate incluing both city and time fixed effects.

We also find that the likelihood of a tweet in Brasília being neutral towards Bolsonaro decreased by 2.14 percentage points, with a significance level of 5% for model (4). It is a larger effect than what we find from the likelihood of opposing Bolsonaro, meaning it is not explained only by an increase in tweets against the far-right. This indicates that there could have been increased political polarization in Brasília.

	Model Specification			
Variable	(1)	(2)	(3)	(4)
Intercept	0.030*	N/A	0.039*	0.044*
	(0.005)		(0.004)	(0.001)
Treated	0.026*	N/A	0.005	N/A
	(0.005)		(0.045)	
Time_Treated	0.019*	N/A	N/A	N/A
	(0.006)			
Treated * Time_Treated	0.011†	N/A	0.011†	0.011
	(0.006)		(0.007)	(0.007)
Observations	55	N/A	55	55
$\bar{R}^2$	0.458	N/A	0.302	0.126
F	14.370*	N/A	21.069*	2.728

Table 2: Effects on political share. Dynamic Panel from the 02/01/2023 to the 12/01/2013; note that we cut off the last the two days to prevent the ebbing off. Note that if an entry has N/A, we couldn't estimate the model due to collinearity.
* indicates significance at a 5% level, † indicates significance at a 10% level

	Model Specification			
Variable	(1)	(2)	(3)	(4)
Intercept	0.2114*	N/A	0.2212*	0.2219*
	(0.0028)		(0.0009)	(0.0003)
Treated	0.0033	N/A	0.0033	N/A
	(0.0063)		(0.0027)	
Time_Treated	0.0217*	N/A	N/A	N/A
	(0.0042)			
Treated * Time_Treated	0.0147	N/A	0.0147*	0.0147*
	(0.0093)		(0.0041)	(0.0029)
Observations	55	N/A	55	55
$\bar{R}^2$	0.5003	N/A	0.4730	0.2508
F	17.024*	N/A	18.852*	13.054*

Table 3: Effects on likelihood of being a tweet critical of Bolsonaro. Dynamic Panel from the 02/01/2023 to the 12/01/2013; note that we cut off the last the two days to prevent the ebbing off. Note that if an entry has N/A, we couldn't estimate the model due to collinearity.
* indicates significance at a 5% level, † indicates significance at a 10% level

Variable	Model Specification (1)	(2)	(3)	(4)
Intercept	0.6519*	N/A	0.6396*	0.6390*
	(0.0033)		(0.0010)	(0.0003)
Treated	-0.0031	N/A	-0.0031	N/A
	(0.0075)		(0.0031)	
Time_Treated	-0.0270*	N/A	N/A	N/A
	(0.0049)			
Treated * Time_Treated	-0.0214†	N/A	-0.0214*	-0.0214*
	(0.0111)		(0.0045)	(0.0028)
Observations	55	N/A	55	55
$\bar{R}^2$	0.5373	N/A	0.5639	0.3931
F	19.743*	N/A	27.151*	25.264*

Table 4: Effects on likelihood of being a tweet neutral of Bolsonaro. Dynamic Panel from the 02/01/2023 to the 12/01/2013; note that we cut off the last the two days to prevent the ebbing off. Note that if an entry has N/A, we couldn't estimate the model due to collinearity.
* indicates significance at a 5% level, † indicates significance at a 10% level

5 Conclusion and Future Research

To the best of our knowledge we are the first ones to investigate the effect of the January 8th insurrection in Brazil. This paper is one of few that attempts a sentiment analysis using advanced NLP-techniques in the Portuguese language. We highlighted the challenges of pre-processing and working with a large amount of non-English tweets (and text, generally). Furthermore we provided alternative methods to traditional techniques.

Our analysis reveals a 1% increase in politicised tweets in Brasília, with a significance level of 10%. However, this effect may be underestimated due to two factors: (1) the share of political tweets in the LDA sample (11.5%) is much lower than the entire sample (40%[1]). (2) The nature of Twitter data is inherently volatile, so the effect could dissipate over time. We also found an increase in the likelihood of tweets being opposed to Bolsonaro and a decrease in the probability of neutral tweets in Brasília.

An extension of this paper would gather more control cities in order to create a synthetic control for Brasília. Additionally, from the results shown in Table 4, we would like to be able to estimate the effect of the insurrection on polarization, as there is some indication that the probability of tweets being supportive of Bolsonaro could have also increased. However, as previously mentioned, we did not have parallel trends in this outcome to justify the use of the difference-in-differences estimator.

Finally, an alternative extension of this paper could be attempting to predict participation in the insurrection itself, something that has been done with Asur and Huberman (2010) and Bollen et al (2011). In the lead up to the invasion, dozens of buses transported supporters from all over the country to the capital. We would like to investigate the difference in sentiment between what we call radical Bolsonaro supporters, who participated in the invasion, and the non-radical Bolsonaro supporters, who did not participate.

[1]We checked the occurrence of our prior words on the full sample

6 References

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (Vol. 1, pp.492-499). IE

Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 5, No. 1, pp. 450-453).

Camara, E. M., Valdivia, M. M., López, L., & Montejo-Ráez, A. (2014). Sentiment analysis in Twitter. Natural Language Engineering, 20(1), 1-28.
https://doi.org/10.1017/S1351324912000332

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. Processing, 150.

Harting, H., & Dothery, C. (2021, September 2). Two Decades Later, the Enduring Legacy of 9/11. PewResearch Center. https://www.pewresearch.org/politics/2021/09/02/two-decades-later-the-enduring-legacy-of-9-11/

Hitesh, M. S. R., Vedhosi, V., Abhishek, Y. J. K., Harsha, K. S., & Santoshi, K. (2019). Real-time sentiment analysis of 2019 election tweets using word2vec and random forest model. In 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT) (pp. 146-151). IEEE.

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10) (pp. 1320-1326). European Language Resources Association (ELRA).

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine

Ramasubbareddy, S., Kumar, K., Aravindharamanan, S., & Govinda, K. (2020). Twitter Sentiment Analysis Based on US Presidential Election 2016. In Advances in Computer Communication and Computational Sciences (pp. 363-373). Springer. https://doi.org/10.1007/978-981-13-9282-5_34

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter Events. Journal of the American Society for Information Science and Technology, 62(2), 406-418.
https://doi.org/10.1002/asi.21462