Sebastián Vallejo Vera, Rage Within the Machine: Activation of Racist Content in Social Media. *Latin American Politics and Society* vol. 65, number 1, Spring 2023.

**Online Appendix**

**Appendix A: Detecting Racist Tweets**

Between October 1 and October 24, 2019, I collected three waves of Twitter data using the strings ``paro'' and ``ecuador'', two terms that are politically and racially neutral and were used by the government and indígena supporters alike. To collect this data, I connected *rtweet* (Kearney 2018) to Twitter's backward search application programming interface (API), gathering tweets in duration of the unrest in Ecuador. The data includes 2,425,239 posts by 85,249 unique Twitter users for the Ecuadorian case.

To detect racist tweets, I adopt a multi-step classification approach, similar to that employed by ElSherif et al. (2018). I start by defining a racist attack towards a member of the indígena community or towards the indígena community in general as a ``negative or hateful comment targeting someone because of their indigenous identity,'' a variation of the definition used by the Google's API *Perspective*.[[1]](#endnote-1) I use Google's API *Perspective*, a content moderating tool that is the industries’ standard for automatic detection of toxic content in written comments. *Perspectiv*  uses a convolutional neural network that scores the likelihood a text contains an identity attack.[[2]](#endnote-2) *Perspective* provides an identity attack score from 0 to 1, interpretable as the probability that a text will be perceived as an identity attack. I use a threshold score of 0.85 to create a dummy for whether a comment is an identity attack or not.

Since we are specifically interested in racist discourse directed towards the indígena community, I create a dictionary with key-phrases that identify the indígena community or members of the indígena community (i.e., *indígena, indio*). I keep only identity attacks (tweets) that contain ``indígena'', the term most used to refer to a member of the indígena community (e.g., ``el/la indígena'') or the indígena community in general (e.g., ``los indígenas''). An alternative, more charged term to refer to indígenas is ``indios.'' ``Indio'' (*Indian*) is often used by the blanco-mestizo population as a derogative identifier, and despite the long history of the indígena community reclaiming the term, it is still widely employed. However, the use of ``indio'' does not automatically reflect racism. Thus, I follow a similar procedure as before (i.e., detect identity attacks and keep those that include the term ``indio''), but lower the threshold score to 0.75, given the charged nature of the term. Figure A.1 provides a scheme of the step-by-step process.



**Figure A.1:** Step-by-step process to detect racist tweets.

Finally, I create a second dictionary with local forms of racist discourse that the algorithm is unable to detect. The following phrases were used: ``indi[ao] de mierda,'' ``long[ao] de mierda,'' ``indi[oa] mmv,'' ``indios\* salvaje,'' ``indios\* apestodo,'' ``indio incivilizado,'' ``emplumado\*,'' ``este indio hijo,'' ``plum[ií]fero,'' ``indi[oa]s mmv,'' ``estos indios hijos de,'' ``[gh]uangudo,'' ``indios\* sucios\*,'' ``regresen al p[áa]ramo,'' ``indios\* feos\*,'' ``indios\* brutos\*,'' ``indio hdlgp,'' ``verga indio,'' ``indio concha de,'' ``indios que son tontos,'' ``indio comido,'' ``indios\* violentos\*,'' ``indios\* delincuentes\*,'' ``indio\* subversivo\*,'' `` long[ao],'' ``long[ao] ,'' ``la indiada,'' ``indios\* tontos\*.'' Additionally, I added tweets when the phrase ``indi[ao]'' appeared in conjunction with one of the following phrases: ``hdgp|hdp|hijo de puta|put[oa],''[[3]](#endnote-3) ``mmv|mamaverga|mama verga|a la verga,'' ``long[oa],'' ``salvaje,'' ``mierda.'' After applying these filters, I identify 1,371 (2%) unique racist tweets in the pro-government community. This multi-step process addresses both aspects of the definition for racist discourse: 1) the *Perspective* algorithm detects ``negative or hateful comment targeting someone because of their identity'', and 2) the key-phrases dictionary identifies the indígena community or an indígena as the target of the toxic tweet.

To check the internal validity of our measure, I hand-annotate a random subset of 1,500 tweets and compare the results to the ones obtained in our procedure.[[4]](#endnote-4) In Table A.3 I report the confusion matrices for the following models, starting with the model used in the main text: (b) indígena \* toxicity threshold < 0.85 + indio \* toxicity threshold < 0.75 + second dictionary; (c) indígena \* toxicity threshold <0.5 + indio \* toxicity threshold < 0.5 + second dictionary; (d) indígena \* toxicity threshold < 0.75 + indio \* toxicity threshold < 0.75 + second dictionary; (e) indígena \* toxicity threshold < 0.85 + indio \* toxicity threshold < 0.85 + second dictionary. Additionally, I include a sample confusion matrix (a).

Diagram

Description automatically generated

**Table A.2:** Confusion matrices for difference racism-detection models

To better interpret the results, in Table A.2 I report a series of measurements derived from the confusion matrix. I base performance of the models on precision (PPV),[[5]](#endnote-5) recall (rec.),[[6]](#endnote-6) $F\_1$ score,[[7]](#endnote-7) and the Matthews correlation coefficient (MCC).[[8]](#endnote-8) Based on this performance metrics, the model used in the main text outperforms other variants. Particularly noteworthy is that the main model predicts true positives at a high rate (recall value of 0.95) without sacrificing precision (0.85)--i.e. without increasing the rate of false positives or Type-I error--.

Table

Description automatically generated

**Table A.2:** Classification results for different multi-step models.

The results from the model not only validate the multi-step process used, but also point to the importance of understanding the contextual nature of certain discursive behavior, such as racist discourse. While the base algorithm of our process, Google's API *Perspective*, does a pretty good job at identifying identity attacks, it does not, and is not expected to, detect the specificities of a language and a region. The attractiveness of the multi-step process used in this paper is that it takes advantage of readily available, open-source, and widely tested tools and modifies them to suit the needs of the researcher. In different geographies than Ecuador, the same logic used in this paper can be applied. For example, if we were to extend the current method to the Bolivian context, we would consider the way language is used to refer to the Bolivian indígena population, both descriptive and derogatory, and create a dictionary based on these. Next, we could create a second dictionary that accounts for local forms of aggression and local racist expressions. We can then run our corpus through *Perspective*'s algorithm, retrieve identity attack scores, and interact them with our dictionaries, as previously explained.

**Appendix B: Robustness checks**

In the main text, the bandwidths for the regression discontinuity models were estimated using a mean square error (MSE) optimal bandwidth selector (Calonico et al. 2020). To ensure the robustness of the findings presented in the main text, I replicate the RDD models using ad-hoc bandwidths. The results show the effect of each event (i.e., Vargas' call and the end of the strike) with intervals of 10 minutes from the event up to two hours before and after the event. Figure B.1 shows that the main results are not driven by the choice of the bandwidth. For all the models on racist tweets we report in Figure B.1, the vast majority of point estimates predict a decrease on time-to-retweet after the event, as suggested by the theory. As reported in the main text, events that threaten the status of the in-group will activate racist content.

1. Varga’s Call



1. End of Strike



**Figure B.1:** Time-to-Retweet during the Ecuadorian protests using local bandwidths. The y-axis plots the robust estimates and 95\% confidence intervals. The x-axis represents the bandwidths in minutes.

For both events, we have direct evidence and can observe the time at which the threat was made, or adjudication was announced. Therefore, concerns over treatment assignment are mitigated. Additionally, since our running variable is time, sorting across the cutoff is not a concern, as in other regression discontinuity models. However, a methodological challenge with time data relates to the validity of the event compared to other similar shocks in our outcome variable over time. To address these concerns, I provide placebo tests.

For this test, we estimate our models for every ten-minute interval over four hours in the pretreatment data. Each interval is considered a placebo cutoff. All models are estimated using the original bandwidths used in the main models: 92 minutes for the Vargas' call, and 73 minutes for the end of the strike. Figure B.2 reports the results of placebo tests. The true treatment affect (red line) falls outside the distribution of the point estimates using the placebo cutoffs. For the Vargas' call, the true effect overlaps with one of the tails of the placebo check; however, the true treatment falls outside the 99% lower bound of the placebo point estimate distribution. Taken together, the placebo checks demonstrate how the treatment effects of the chosen events are *unlikely to be driven by a random unobserved shock in the running variable*.



**Figure B.2:** Placebo checks for the effect of threats to the status of the in-group in time-to-retweet. The red lines represent the time of the event using the true cutoff. The x-axis plots the robust point estimates using placebo cutoffs on every ten minutes.

As pointed out in the main manuscript, one possible threat to the validity of the model is the potential of sorting by users right before the event. In our case, that users will anticipate the outcome of the events (i.e., a perceived threat to the status quo) and change their behavior accordingly. Given that I expect the effects of the event to decrease *time-to-retweet*, any anticipation to the treatment is likely to go in the same direction. Therefore, our model would underestimate the effects of the treatment, suggesting that the true effect of the event would be stronger.

However, I run a McCrary test to evaluate whether there is any evidence of sorting around the cutoff. The initial bandwidth is chosen via procedures described in McCrary (2008). I then follow Harris (2021) and test a variety of smaller (multiple less than one) and larger (multiple greater than zero) bandwidths to test the robustness of the failure to reject the null. Across most specifications (see Tables B.1 to B.4), the test fails to reject the null hypothesis of no discontinuity at the cutoff, in particular for racist tweets. This suggests that sorting around the cutoff is not a concern.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table B.1:** Sorting around the cutoff for racist tweets during Vargas' call: McCrary test using multiple bandwidths fail to reject the null hypothesis of no discontinuity in the density of the running variable. | | | | |
| **Bandwidth (left)** | **Bandwidth (right)** | **Multiple** | **Est. Difference** | **p-value** |
| 92.26 | 141.80 | 0.25 | -0.00 | 0.63 |
| 147.62 | 226.87 | 0.50 | -0.00 | 0.69 |
| 405.94 | 623.90 | 1.00 | -0.00 | 0.18 |
| 461.30 | 708.98 | 1.25 | -0.00 | 0.83 |
| 553.56 | 850.77 | 1.50 | 0.00 | 0.46 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table B.2:** Sorting around the cutoff for non-racist tweets during Vargas' call: McCrary test using multiple bandwidths reject the null hypothesis of no discontinuity in the density of the running variable. | | | | |
| **Bandwidth (left)** | **Bandwidth (right)** | **Multiple** | **Est. Difference** | **p-value** |
| 53.41 | 84.29 | 0.25 | -0.00 | 0.61 |
| 106.83 | 168.58 | 0.50 | -0.00 | 0.89 |
| 213.65 | 337.16 | 1.00 | 0.00 | 0.75 |
| 267.07 | 421.45 | 1.25 | -0.00 | 0.03 |
| 320.48 | 505.73 | 1.50 | -0.00 | 0.00 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table B.3:** Sorting around the cutoff for racist tweets during the end of strike: McCrary test using multiple bandwidths fail to reject the null hypothesis of no discontinuity in the density of the running variable. | | | | |
| **Bandwidth (left)** | **Bandwidth (right)** | **Multiple** | **Est. Difference** | **p-value** |
| 165.57 | 70.73 | 0.25 | -0.00 | 0.39 |
| 331.13 | 141.47 | 0.50 | -0.00 | 0.18 |
| 662.27 | 282.94 | 1.00 | 0.00 | 0.33 |
| 874.19 | 373.48 | 1.31 | 0.00 | 0.06 |
| 993.40 | 424.41 | 1.50 | 0.00 | 0.28 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table B.4:** Sorting around the cutoff for non-racist tweets during the end of strike: McCrary test using multiple bandwidths reject the null hypothesis of no discontinuity in the density of the running variable. | | | | |
| **Bandwidth (left)** | **Bandwidth (right)** | **Multiple** | **Est. Difference** | **p-value** |
| 35.71 | 28.67 | 0.25 | 0.00 | 0.08 |
| 71.41 | 57.34 | 0.50 | 0.00 | 0.85 |
| 142.82 | 114.69 | 1.00 | 0.00 | 0.00 |
| 178.53 | 143.36 | 1.25 | 0.00 | 0.00 |
| 214.24 | 172.03 | 1.50 | 0.00 | 0.00 |

I also test the sensitivity of results to local smoothing near the cutoff. To do so, I run a doughnut analysis on my RD designs. The doughnut analysis tests the sensitivity of the design by removing observations around the cutoff. The results in Figures B.3 and B.4 suggest the results are not driven by observations around the cutoff.

**Table B.3:** Doughnut Estimates for Vargas’ Call (Left: Racist Tweets; Right: Non-racist Tweets)



**Table B.4:** Doughnut Estimates for End of Strike (Left: Racist Tweets; Right: Non-racist Tweets)



Finally, one concern about the research design might be that the salience and information provided by non-racist and racist tweets differ at the time of the event. If racist tweets also provide more salient information than non-racist tweets, salience might be why users are reacting faster rather than the racist frame. Some elements mitigate this concern: 1) the initial selection of tweets (during the scraping stage) already limits content related to the strike, and 2) highly salient events concentrate the attention of users (Lin et al. 2013), and the reduced period analyzed in the RD would suggest that all users are talking about the event (while providing different frames). A revision of a selection of tweets close to the cutoff suggests that the opposite might be happening (i.e., racist tweets seem to be less informative than non-racist tweets). Racist tweets were limited to attacking the indígena community or leaders, while the context is inferred from events occurring when users produced the tweets. For example, one of the less aggressive racist tweets would read: “trabaje 20 años con indígenas en todos esos años nunca conocí uno de ellos uno que sea con visión de futuro [sic] son demasiados egoístas ignorantes racistas nunca te aceptan una recomendación le compadezco al mediar con estas bestias es algo casi imposible”. This translates to: “For 20 years I worked with indígenas and in all those years I have never met one of them with foresight. They are too selfish, ignorant, and racist. They never accept any advice. I pity having to mediate with these beasts, it is impossible". The user refers to the UN envoy's mediation to end the strike. Non-racist content was more informative as users retweeted news outlets reporting the events. However, many non-racist retweets were as uninformative as the racist tweets, but without the racist language.

**Appendix C: Diffusion of racist content.**

The diffusion of a message depends on many elements, not only agreement with the content: the number of people connected at the time of the tweet was produced, how much of a specific frame is produced, who is producing the content, etc. Thus, testing the total amount of retweets of a specific frame is not only capturing agreement with the content, but other elements as well. However, I also test if threats to the in-group increased the probability of a retweeted post being racist. The results are presented in Table C.1.

**Table 2:** Regression discontinuity estimation for probability of a racist tweet

|  |  |  |
| --- | --- | --- |
|  | Vargas’ Call | End of Strike |
|  | Model 1 | Model 2 |
|  | Non-racist Tweets | Non-racist Tweets |
| Treatment Effect | 0.022\*\* | 0.051 |
|  | [0.008 , 0.036] | [-0.008 , 0.111] |
| Number of Obs. | 21684 | 13315 |
| Pre-Treatment Obs. | 10852 | 4899 |
| Post-Treatment Obs. | 10832 | 8416 |
| Note: 95% confidence intervals reported in brackets, with confidence levels reported as follows: \*p < .1; \*\*p < .05; \*\*\*p < .01. Model 1 centers treatment on October 10, 2019, at 11:51 AM, local time, when Jaime Vargas, the president of the CONAIE, asks the Police force and the military to join the protests and disobey the orders of the government. Model 2 centers treatment on October 11, 2019, at 9:45 PM, local time, when the United Nations mediator announced the terms agreed upon for the end of the strike. | | |

The results are in line to our initial prediction: threats to the status quo increase the probability of a user retweeting racist content. The increases in diffusion of racist content are more clearly seen after Vargas’ call (p < 0.05), even though the effect if bigger after the end of the strike. At the time of Vargas’ call, the probability of a user retweeting racist content increased by two percentage points. For the end of the strike, the increase was of five percentage points, even though the effect is not statistically significant at the 95% level. I present a visualization of the results in Figure C.1. As previously suggested, the results are noisier, as the number of retweets is influenced by different elements. However, there is some evidence pointing to increases in diffusion of racist content when the status of the in-group is threatened.

Chart

Description automatically generated

**Figure C.1.:** Probability of a racist tweet during the Ecuadorian protests.

Left panel: Centering on October 10, 2019, at 11:51 AM, local time, when Jaime Vargas, the president of the CONAIE, asks the Police force and the military to join the protests and disobey the orders of the government. Right panels: Centering on October 11, 2019, at 9:45 PM, local time, when the United Nations mediator announced the terms agreed upon for the end of the strike.

**Notes**

1. Google's API *Perspective* considers an identity attack any ``negative or hateful comment targeting someone because of their identity.'' [↑](#endnote-ref-1)
2. The model was built using millions of comments from the internet, using human coders to rate the comments on a scale from “very toxic” to “very healthy”, and using this large corpus as training data for the machine learning algorithm. See Wulczyn et al. (2017) for a comprehensive discussion on *Perspective*. [↑](#endnote-ref-2)
3. These are all variations of the same phrase. [↑](#endnote-ref-3)
4. I over sampled predicted racist tweets from the multi-step process to account for the sparsity of racists tweets. [↑](#endnote-ref-4)
5. Precision, or positive predictive value (PPV), is estimated as follows: PPV=TP/TP + FP, where TP is the True Positive, and FP is the False Positive. [↑](#endnote-ref-5)
6. Recall, also known as sensitivity or the true positive rate (TPR), is estimated as follows: TPR=TP/P, where TP is the True Positive, and P is the total Positive. [↑](#endnote-ref-6)
7. The F\_1 score is the harmonic mean of precision and recall. F\_1 score measure test accuracy. [↑](#endnote-ref-7)
8. The Matthews correlation coefficient (MCC) considers the true and false positives, and true and false negatives. The MCC returns a value between -1 and +1, where -1 is no correlation between prediction and observation, 0 represents better than random, and +1 is a perfect prediction. [↑](#endnote-ref-8)