Rage in the Machine: Activation of Racist Content in Social Media

Sebastián Vallejo Vera

ABSTRACT

Racism in social media is ubiquitous, persisting online in ways unique to the internet while also reverberating from the world offline. When will racist frames activate in social media networks? This article argues that social media users engage with racist content when they perceive a threat to the in-group status, selecting frames that serve as markers to separate the in-group identity from the out-group identity. Racialized frames serve as these markers, and the perceived threats to the in-group status make racist content cognitively congruent. Evidence of this behavior is provided by examining Twitter activity during the indígena protests in Ecuador in October 2019. A novel, multistep machine-learning process detects racist tweets, and an interrupted time series analysis shows how events that can be perceived as threats to the in-group activate racist content in some social media communities.

Keywords: Racism, social media, indígena protest, machine learning, Ecuador

Racism in social media is ubiquitous, persisting online in ways unique to the internet while also reverberating from the world offline (Daniels 2013). This article explores the framing of racism in social media, paying particular attention to how individual users—rather than the media or institutions—decide to share content and what motivates users to promulgate these messages throughout their networks. It finds that perceived threats to the status of the in-group play an important role in the acceptance of racist content and the speed of propagation of racist social media posts.

Social media are a powerful tool to deliver and frame political narratives among voters (Barberá et al. 2015; Neumayer et al. 2016; Aruguete and Calvo 2018). In large part, users' self-selection will shape these frames. Social media users will interact with and cluster around like-minded individuals (Himelboim et al. 2013)because the platform subtly encourages these interactions, thanks to the development of sophisticated algorithms. Consequently, social media users will frame events by collectively selecting or discarding content that is then impressed on the walls of like-minded peers. Within these social media bubbles, users accept and share

Sebastián Vallejo Vera is a profesor investigador in the Escuela de Ciencias Sociales y Gobierno, Tecnologico de Monterrey, Monterrey, Mexico. svallejovera@tec.mx. ORCID: 0000-0002-5848-7400. Conflicts of Interest: The author declares that there are no conflicts of interest.

DOI 10.1017/lap.2022.25

frames they find cognitively congruent and discard frames they find cognitively dissonant (Aruguete and Calvo 2018).

These homogenous communities, resembling in-groups—users identifying as members of social groups with shared identities (Keipi et al. 2017)—are exposed to a marketplace of frames. One of these frames is racist content. Research has shown that in political contexts, threats to the in-party identity (Amira et al. 2021) or status (Mason 2016) are drivers of out-group hate and anger. This study extends this logic to the role of racism in social media. I argue that in these homogenous communities of like-minded peers, a perceived threat to the in-group status will activate racist behavior. Racist frames are particularly salient because they appeal to the in-group's identity. When the out-group threatens the status of the in-group, users will select frames that serve as markers to separate the in-group identity more starkly from the out-group identity. Racialized frames serve as these markers, and the perceived threat to the in-group status makes racist content cognitively congruent.

Evidence of this behavior is provided by examining social media activity during the *indígena* (indigenous*)* protests in Ecuador in October 2019, a political crisis triggered by the decision of President Lenín Moreno's administration to eliminate gas subsidies. This study builds a network based on user interaction (tweets and retweets) and identifies various communities relevant to the event studied. It finds three clusters of users, a progovernment community, a pro-indígena community, and a second opposition community.

Focusing mainly on the progovernment community (where all the overt racist content was identified), the study first evaluates the general attitude of users toward racist content. To do this, it implements a survival model to estimate the users' reaction time (retweet) to racist tweets and compare this to the reaction time to nonracist tweets. The findings show that racist content in social media networks is uncommon, yet not rare. Despite the presence of racist messages during the span of the indígena protest, racist frames did not consistently activate the progovernment community (i.e., users reacted faster to nonracist tweets than to racist tweets). Users were unwilling to accept and engage with socially punishable behavior, such as racist content.

Then, to analyze the reaction of users to perceived threats to the status of the in-group, the study looks at two significant events that developed during the strike: the moment the leader of the indígena community, Jaime Vargas, called on police and military forces to disobey government orders and join the protests; and the point when the Moreno government announced the end of the strike after agreeing to the demands of the indígena leadership. By comparing the reaction of users to content around the time of these (publicly broadcasted) events, the study finds that racist frames are more readily accepted and shared faster by progovernment community users when there is a perceived threat to their in-group.[1] To test this, a regression discontinuity design is presented, with *time-to-retweet* as the dependent variable.

Ecuador is an interesting test case to explore racism in social media, a country with an organized and politically active indígena community subject to historical

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

manifestations of marginalization and racism. As a collective, the indígena community has challenged political power and gained political spaces (Van Cott 2008; Becker 2010) yet remains marginalized in the racially stratified Ecuadorian society (Hall and Patrinos 2004). The indígena political mobilization in an exclusionary state led to clearly defined communities with conflicting interests and power relations (see Bretón and Pascual 2003), a reality manifested in its online social networks.

This study also presents a methodological contribution: an easy-to-implement strategy to detect racist tweets. Given the highly contextual nature of racist expressions, current dictionary-based and machine-learning techniques for detecting racism on the web perform poorly when applied to data in other languages or from different geographies. This study uses dictionary-based and semisupervised machine-learning techniques (i.e., Google's Perspective algorithm) to detect racism in the Ecuadorian network. The article explains how to implement this approach in other contexts and discusses the scope and limitations of this approach.

This article begins by examining racism and social media framing to unpack the conditions limiting and heightening the proliferation of racist content on social media. It then introduces the particularities of the Ecuadorian case and presents the Ecuadorian Twitter data and the multistep process employed to detect racism. It uses these data to show how users engage with racist content and the effect of events threatening the in-group status. The article concludes by discussing the argument's implications for the general study of racism in and beyond social media.

Social Media Framing and Racism

In social media networks, users tend to cluster around like-minded peers, which Himelboim et al. (2013) describe as selective exposure. Selective exposure occurs when individuals actively seek information that matches their beliefs, connecting with content that is cognitively congruent with their preferences and prior beliefs. Within these social media bubbles, individuals are exposed to information consistent with their beliefs while deciding what content to accept and share and what content not to accept and consequently not to share. In other words, users are selectively exposed to information cognitively congruent or cognitively dissonant with their preferences and then decide whether to propagate this content across the network (Aruguete and Calvo 2018).

These social media bubbles are a marketplace of noncompeting frames. Given the vertical configuration of some social media networks, it is usually high-degree network authorities, users with a significant number of followers, who are interested in framing social media events to their advantage. Users are exposed to numerous frames and "vote" among the choices. Cognitively congruent frames—that is, frames accepted by users—will propagate; cognitively dissonant frames will go unshared and unseen (Aruguete and Calvo 2018).

Racist content is among the frames users are exposed to. Like other frames in social media bubbles, racist frames will sometimes be cognitively dissonant to users

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

and therefore go unshared; other times, they will be cognitively congruent to users and will propagate. Unlike other frames, racist discourse is socially punished, particularly in public settings (Bonilla-Silva 2015). In other words, racist discourse is a socially costly behavior among a menu of other less socially costly options. As such, people are unlikely to engage without a trigger. When will racist frames, a socially costly behavior, activate social media networks?

Homogenous communities formed in social media, particularly those in polarized environments, resemble in-groups with shared identities and social homophily (Keipi et al. 2017).[2] Social affiliations to gender, religious, and ethnic or racial groups promote in-group bias: greater attachments to and preference for in-group members (Tajfel 1981). For example, in political parties, these affiliations motivate members to advance the party's status (Huddy 2001).

However, in-group love is not reciprocal to out-group hate (Brewer 1999). Desires to benefit the in-group (or the in-party) do not necessarily drive biased behavior toward out-group members (or out-party members). Denigrating the out-group does not advance the in-group or the in-party status. Instead, in political contexts, threats to the in-party identity (Amira et al. 2021) or status (Mason 2016) are drivers of out-group hate and anger. Furthermore, research has shown that anger is a powerful political mobilizer (Groenendyk and Banks 2014), especially in strong partisans (Huddy et al. 2015).

I extend this logic to the role of racism in social media. I argue that in these homogenous communities of like-minded peers, threats to the in-group status will activate racist behavior. Racist frames are particularly salient because they appeal to the in-group's identity. When the out-group threatens the status of the in-group, users will accept and propagate faster frames that serve as markers to separate the in-group identity more starkly from the identity of the out-group. Racialized frames serve as these markers, and the threat to the in-group status makes racist content cognitively congruent. The mobilization of in-group users in social media is carried out by engaging with and propagating content (e.g., racist content), which occurs with cognitively congruent frames. From the proposed theory, the formulation of the hypotheses follows:

Hypothesis 1. *Overall, users will reject racist frames and decrease the speed at which they share racist content.*

Hypothesis 2. *Events where the status or identity of the in-group is perceived to be threatened will activate racist frames among users.*

The findings from Amira and colleagues (2021), Mason (2016), and Groenendyk and Banks (2014) are relevant in explaining the activation of racist frames in social media, especially when users frame political events. Social media have become a battleground where political narratives are delivered and framed among voters (Barberá et al. 2015; Neumayer et al. 2016; Aruguete and Calvo 2018). The political network communities created by the selective exposure and dissemination of content often align with the political camps contending for political power in

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

"real life." Thus political psychology and political communication literature insights inform our understanding of users' engagement with racist content in social media. Mainly, users will engage with racist content, one of many frames shared in social media bubbles, when the status of the in-group is perceived to be threatened. Consequently, racialized frames, especially those attacking the out-group, will be more readily accepted and shared faster by users.

Racist Discourse in Social Media

The literature on Latin America (Martínez-Echázabal 1998) and Ecuador (Roitman 2009) notes that race is a complex construction, due to *mestizaje* and the strong correlation between ethnic background, perceptions of race, and class. In other words, how issues of race are framed, and how race is socially constructed, have geographical caveats. In Ecuador, as in other parts of the world, discursive racism is usually framed as intolerance toward the out-group (e.g., indígena) by creating strict demarcations between the self and the "other." However, individuals often dismiss racism and race as explanations for racist behavior, and state structures and government representatives often pay mere lip service to integration and ethnic identity. It is not surprising that racist language is normalized (Roitman and Oviedo 2017), even though this characteristic has different degrees and forms (De la Torre 1996). Everyday patterns of behavior and speech and the organization of the state are configured in a way that "indios" and "indígenas" are the subjects and objects of structural discrimination.

We would expect online social spaces, such as Twitter, to replicate public and private racial discourses, especially since online spaces amplify racist discourse and unmask where racist discourse is produced and how it is reproduced (Eschmann 2020). The internet is a hybrid social space, at once public and private (Daniels 2013), where established and new forms of racism are facilitated (Daniels 2013; Nakamura 2007). The user-generated communities on online platforms encourage intimate discursive interaction predicated on racial identity (Brock 2009; Daniels 2013).

To a varying degree, social media platforms allow for technical anonymity and social anonymity. Some studies suggest that this can explain aggressive and hostile online behavior, given a user's perceived freedom from social standards and sanctions (Christopherson 2007; Lapidot-Lefler and Barak 2012), even though there is also evidence to the contrary (Jaidka et al. 2022). However, not all platforms grant or encourage the same levels of anonymity (Barlett et al. 2018). Given the importance of the relational component in less anonymous social media networks such as Twitter (Preussler and Kerres 2010), users will be cognizant and cautious of the real-world ramifications of their actions (Barlett et al. 2018).

Racism has diverse discursive manifestations, both overt and covert.[3] From a practical standpoint, this study focuses on overt forms of racist content, which are easier to detect systematically than other forms of racism; in Ecuador, overt racist forms explicitly target someone because of their indigenous identity, using negative

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

and hurtful comments. Most important, in a country such as Ecuador, many platforms and institutions, classes, and contexts reproduce a "blanco-mestizo" racist ideology in often subtle and normalized patterns that the user is not aware of or would not consider racist (Roitman and Oviedo 2017). Overt racism dispels any ambiguity.

The #ParoEcuador Context

On October 1, 2019, Lenín Moreno, then president of Ecuador, announced the elimination of gas subsidies. Two days later, the *Unión de Transportistas* (Transportation Union) announced a strike.[4] In addition to the transportation sector, eliminating gas subsidies had a dire effect on the indígena community. Increases in gas prices led to increases in the prices of goods, ultimately affecting the poorer sectors of Ecuadorian society, disproportionally represented by the indígena population. Two days after the Unión de Transportistas announced its strike, the Confederation of Indigenous Nations of Ecuador (CONAIE), the largest indígena organization in the country, followed suit, announced a national strike, and started mobilizing its base to move toward the capital, Quito. The indígena mobilization was a reaction to the consequences of eliminating the gas subsidies and a continuation and expansion of ongoing local protests about the general lack of government support for the community's problems.

By the time the indígena movement reached the capital, the president had moved the seat of government to Guayaquil and declared a state of emergency. In a polarized environment, pro- and antigovernment media and protestors displayed contrasting sentiments toward the actions of the executive and the protesters and widely differing accounts of violent incidents. The political environment was complex, as the indígena-led protests incorporated an ample array of social and political actors, among them the *Frente Unitario de Trabajadores* and *Frente Popular* (two workers' unions) and various student movements. Furthermore, followers of former president Rafael Correa joined the protest against the government. While indígena leaders consistently disassociated themselves from Correa and Correa supporters who also took to the streets, the government and some of its supporters blamed the strike on Correa. For example, some of the content reproduced in social media framed the indígena movement as being manipulated by "useful idiots" (*tontos útiles*).

Posts related to the protests circulated extensively on Facebook and Twitter, the social media outlets with the largest user bases in Ecuador (AmericasBarometer 2018).[5] As state violence increased, many of the reports were initially broadcast by online media sources before being picked up by the more traditional outlets. The country was paralyzed for more than ten days, prompting mixed reactions from different groups. Labor unions joined the protests and various organizations, including universities, supported the indígena movement as government violence increased. However, the protests also paralyzed an already faltering economy. Business representatives, for example, denounced the indígena protest and praised the government for its position (Díaz and Mejía Artieda 2020).

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

Beyond the role of social media in the Ecuadorian protest, there are structural characteristics worth discussing. Racism in Ecuador is a “system of ethnic-racial dominance” historically rooted in European colonialism (Van Dijk 2005), directed, in great part, toward the indígena population (Beck et al. 2011). The state has marginalized the indígena population and done little to support its communities or grant them equal access to political spaces. Regardless, the indígena population has managed to organize around a unified banner (i.e., the “indígena” banner, despite the different and sometimes conflicting nationalities) to demand and achieve important political and social victories (see Becker 2010). However, these victories have done little to change the racist ideology that permeates all levels of the state and society. Despite their political activism and mobilization, indígenas and the indígena community are still marginalized and remain the main target of the national racist ideology.

The #ParoEcuador Data

Between October 1 and October 24, 2019, I collected three waves of Twitter data using the strings *paro* and *Ecuador*, two politically and racially neutral terms used by the government and indígena supporters alike. To collect these data, I connected *rtweet* (Kearney 2019) to Twitter’s backward search application programming interface (API) and gathered tweets for the duration of the unrest in Ecuador. The data include 2,425,239 posts by 85,249 unique Twitter users for the Ecuadorian case. Of this sample, 93 percent were retweets of an original tweet.

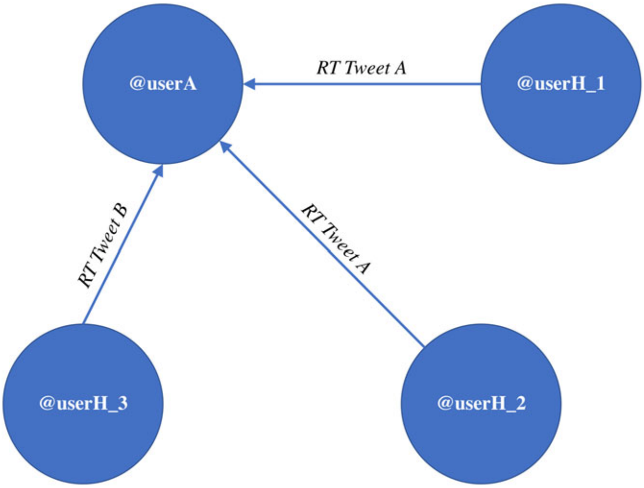
I restricted data collection to various days before the beginning of the indígena strike and various days after the end of the strike. By limiting the data to October 2019, I could be more confident that the dialogue between users was about the actions of the indígena community during the strike. Throughout, the data show a latent level of perceived threat to the in-group status. Before the national conversation centered on the events surrounding the strike, little racist content was produced because little attention was paid to the indígena community.

The #ParoEcuador Network

Selective exposure in social media leads to homogenous and differentiated communities. To estimate these communities in the Twitter data, I first built a directed network in which each user was a node and an edge was created when a user *H* (hub) retweeted user *A* (authority).[6] To illustrate this interaction, figure 1 shows a diagram of how the Twitter network is created.

Communities—clusters of nodes where the same information (tweets) is shared—are identified via a random-walk community detection algorithm (Pons and Latapy 2005).[7] This algorithm identified three primary communities in Ecuador: a progovernment network, which included 36,579 nodes; an indígena community network of 29,624 nodes; and a pro-Correa community network of 25,376 nodes. These communities are in line with the main groups disputing in

Figure 1. The Formation of a Network in Twitter



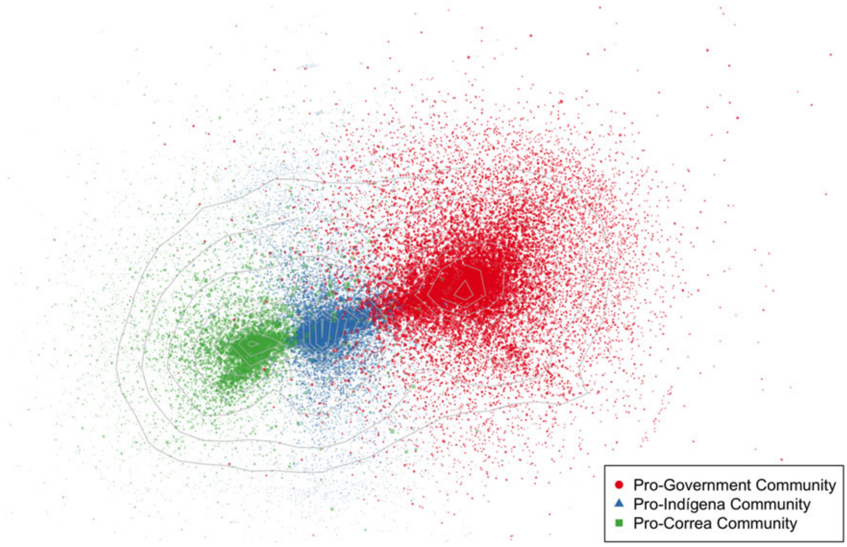
Notes: In the figure, @userA tweeted two tweets: "Tweet A" and "Tweet B." @userH_1 and @userH_2 retweeted "Tweet A," while @userH_3 retweeted "Tweet B." Each user is a node in the network. When a user (node) retweets another user (node), a link (edge) is created between them. In the diagram, each arrow is an edge. Since @userH_1 retweeted a tweet from @userA, @userH_1 is called a hub, and @userA is the authority. Users (nodes) can be both hubs (when they retweet other users) and authorities (when other users retweet them).

the political battle during the protests (Díaz and Mejía Artieda 2020). Figure 2 presents a basic Fruchterman-Reingold layout of the Ecuadorian network.[8] It describes the progovernment community with blue squares, the indígena community with red triangles, and the pro-Correa community with green circles. The size of the nodes is proportional to the nodes' in-degree, with larger nodes indicating users retweeted by a larger number of followers.

On Twitter, communities formed around political events and cleavages often have at their center political leaders or users strongly aligned to the leadership. Influential authorities in the 2016 US election communities included presidential candidates @HillaryClinton and @realDonaldTrump; for the 2018 #Tarifazo networks in Argentina, it was the opposition leader @CFKArgentina (Cristina Fernández de Kirchner) and then-president @mauriciomacr (Mauricio Macri). In the Ecuadorian network during the 2019 indígena protests, the progovernment community had at its center President @lenin, Vice President @ottosonnenh, and interior minister @mariapaularomo, as well as other prominent progovernment users. In the center of the pro-indígena community was the institutional account of @CONAIE_Ecuador and its president, @jaimevargasnae. And in the center of the pro-Correa community were former president @mashirafael and high-ranking members ofhis party. Beyond public officials or politicians, other influential users included media personalities, media outlets, and social media commentators.

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

Figure 2. Primary Connected Network During the Ecuadorian Protests



Notes: Ecuadorian protests between October 1 and October 24, 2019. Red triangles describe progovernment users. Blue squares describe users aligned with the indígena community. Green circles describe pro-Correa users.

The difference and sparsity of exchanges between each community give an account of the polarization of the Ecuadorian network. Of all edges in the indígena community, 78.0 percent are with members of the same community (i.e., indígena community → indígena community), and only 6.5 percent with members of the progovernment community.[9] Of all edges in the progovernment community, 91.6 percent are with members of the same community (i.e., progovernment community → progovernment community), and only 4.4 percent with members of the indígena community.

It is important to note that Twitter is not a representative sample of the Ecuadorian population. According to LAPOP's AmericasBarometer project (2018), social media users in Ecuador are more likely to be young, urban, educated, and well-off. Despite these shortcomings, using Twitter data has several advantages. Unlike other, more popular social media sites, Twitter data are readily available. Twitter's hierarchical nature resembles the political dynamics of real life, as a few important figures produce most of the information. At the same time, many peripheral users decide whether or not to replicate it (while not producing much information on their own). Furthermore, although they do not reflect the general population, Twitter networks, including the Ecuadorian network, are consistent and stable across time (Calvo and Aruguete 2020) and reflect the similar polarized interactions witnessed on the streets and in the press.

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

Detecting Racist Tweets

A large body of research is dedicated to detecting hate speech on social media (Schmidt and Weigand 2017; ElSherief et al. 2018; Chatzakou et al. 2017), but accurate and systematic hate speech detection is a challenging task (Davidson et al. 2017). Despite the many advances on the topic, there are still limitations to the automatic detection of racist discourse. Not the least of these is the contextual nature of racism (Van Dijk 2005). While racist discourse will work to maintain existing power structures and racist ideologies, the language will evolve and shift, depending on the particularities of society and time. Thus, even if we were to take pretrained models to detect hate speech, these would be useless in contexts different from the ones they were originally trained for.

To solve this problem, I adopted a multistep classification approach, like that employed by ElSherief et al. (2018). I started by defining a racist attack toward a member of the indígena community or toward 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*.[10] I used Google's API *Perspective*, a content-moderating tool that is the industry standard for automatic detection of toxic content in written comments. *Perspective* uses a convolutional neural network that scores the likelihood a text contains an identity attack.[11] *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 used a threshold score of 0.85 to create a dummy for whether a comment is an identity attack or not.[12] The *Perspective* algorithm was trained to detect identity attacks on frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Since this study is specifically interested in racist discourse directed toward the indígena community in Ecuador, I created a dictionary with key phrases that identify the indígena community or members of the indígena community (i.e., *indígena, indio*). I kept only identity attacks (tweets) that contained *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*. The blanco-mestizo population often uses *indio* (indian) as a derogative identifier. Despite the indígena community's long history of reclaiming the term, it is still used to signal a racial attack. Therefore I followed a similar procedure as before (i.e., detect identity attacks and keep those that include the term *indio*) but lowered the threshold score to 0.75, given the charged nature of the term. Even though the term *indio* does not automatically reflect racism, it increases the probability of a text being racist (primarily when someone outside of the indígena community uses it). In addition, I created a second dictionary with local forms of racist discourse that the algorithm cannot detect. For example, during the protest in Ecuador, the phrase *emplumados* (feathered), was used mockingly to describe indígena leaders.[13]

To check the internal validity of the measure, I hand-annotated a random subset of 1,500 tweets and compared the results to the ones obtained in the main procedure.[14] In a sample limited to tweets from the progovernment community, the multistep model obtained an F1 score of 0.89, with a recall score of 0.85 and a precision score of 0.94, suggesting that the model is accurate and particularly good at avoiding false positives (Type-I error).[15] No overt forms of racism were found in tweets produced by users from the indígena or pro-Correa communities in the hand-annotated sample. Nevertheless, the multistep process detected racist tweets in those communities, a limitation discussed below. After applying these filters, I identified 1,371 (2 percent) unique racist tweets in the progovernment community. This multistep process addresses both aspects of the definition of racist discourse: 1) the *Perspective* algorithm detects "negative or hateful comment targeting someone because of their identity," and the key phrase dictionary identifies the indígena community or an individual indígena as the target of the toxic tweet.

This multistep process has some limitations. The first and most noteworthy is its reliance on the *Perspective* algorithm to identify racist tweets. Hosseini et al. (2017) show drawbacks to the *Perspective* toxic detection system, mainly underdetecting toxicity when keywords (e.g., words that signal toxicity) are misspelled and sending false alarms for benign phrases denouncing toxic behavior.[16] The latter was particularly problematic for tweets produced by users from the indígena or pro-Correa communities. While the rate of racist tweets detected by the model in these communities was low relative to those identified in the progovernment community, all of these were incorrectly identified as being racist.[17] These tweets usually denounce racist behavior and discourse from other users, political figures, media, and the police or point out racism in another user's tweet. Therefore I focused solely on users from the progovernment community.

A second limitation is that the *Perspective* algorithm cannot capture subtler forms of racist discourse, such as those that infantilize indígenas (Guerrero 1997) or use irony or wordplay to mock indígenas (Sue and Golash Boza 2013). These forms of racism can be found in all the communities, including the indígena and opposition communities. However, because of their subtlety, I could not systematically detect them through the model, so I limited the analysis to overt forms of racism. Exploring subtler forms of racism is a promising avenue for further work.

In addition, Twitter has various mechanisms that detect and potentially eliminate tweets with racist content.[18] Unfortunately, it is impossible to know whether a tweet was deleted or if it was deleted before the collection of the corpus (see Timoneda 2018).[19] Overall, most of the limitations led to underreporting of racist discourse in the corpus. Underreporting leads to Type II error, with a higher probability that the effects of racist discourse are underestimated. Thus, in a worst-case scenario, my findings are conservative estimates of the true effect.[20]

User Activation and Variables of Interest

The primary variable of interest is whether a tweet is racist or not. We want to know how these messages are accepted and spread across the network. In social media, "acceptance equals propagation" (Aruguete and Calvo 2018). Aruguete and Calvo (2018) show that reaction time is associated with the diffusion of content (i.e., less time, more diffusion). They further show that *time-to-retweet*, a measure of reaction time, is a proxy for latency, which researchers frequently use in experimental settings to measure cognitive congruence or dissonance. Content that is accepted (i.e., cognitively congruent) will propagate. As Aruguete and Calvo (2018, page 12) point out, "acceptance leads to propagation … as users share information, new users are exposed to the media content of their peers."

To operationalize "acceptance" and the speed at which frames are spread in a network, I tested whether latency increased or decreased racist content and how latency for racist content varied under different conditions. Latency is measured as *time-to-retweet*, the number of seconds elapsed from when a user (authority) posts a tweet to when a second user (hub) retweets the same post. Users consuming messages on Twitter will propagate content from their community peers more quickly when these are accepted.[21]

In addition to the tweet's text, the tweet's time, and the retweet time, I collected information on the users' screen name, follower count, friend count, and the status of the users' accounts (verified or not verified). From the network, I computed in-degree —the number of times a user had been retweeted—to identify high-authority users (i.e., users with a significant number of followers) and low-authority users. Users who are more central to the network usually behave differently from less prominent users (Calvo and Aruguete 2020), and this information allows us to control for that effect. I also controlled for the effect of bots by including the iratio of followers to friends.[22] In line with prior research, the Twitter data showed high degrees of concentration. In the #ParoEcuador network, less than 5.6 percent of accounts w responsible for 45 percent of the content circulated in the network.

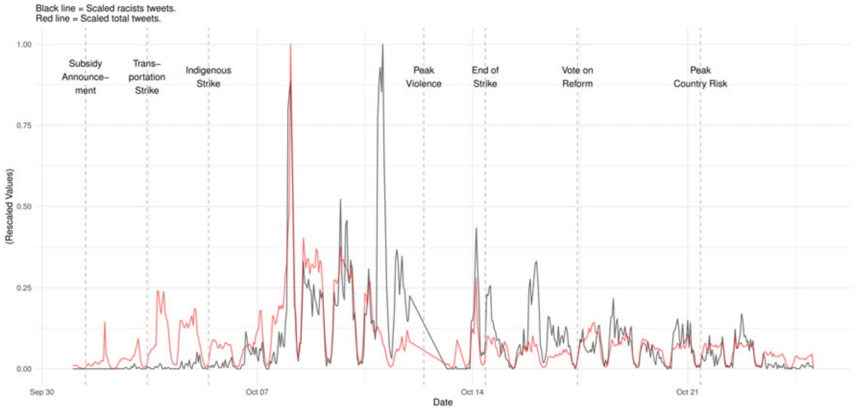
Sharing Racist Tweets in the #ParoEcuador Network

Before analyzing the (racist) reaction of progovernment users to threats to the in-group status, let us look at a more general snapshot of users interacting with racist content. Figure 3 describes user activity of the progovernment #ParoEcuador Network across time, labeled by critical dates The red line represents overall tweets, and the black line racist tweets. The network activity increased right after the Unión de Transportistas announced its strike. However, it was only after the CONAIE announced its strike that the racist content began and continued for some time after the end of the strike, eventually waning, as did the attention to the issue. The timeline shows how users became more active in the network as the strike intensified and how racist activity did not necessarily follow the same pattern. In general, racist content was a latent frame in the progovernment community.

Hypothesis 1 argues that users will reject racist frames and decrease the speed with which they share racist content. To systematically explore how racist language increases (or decreases) latency in Twitter user engagement, the determinants for *time-to-retweet* were analyzed. Higher values for *time-to-retweet* mean longer times

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

Figure 3. Timeline of #ParoEcuador Activity, October 1–October 24, 2019



between the original post of the tweet by the authority and the retweet of the post by the hub. I estimated a Cox proportional-hazards model, with unstandardized coefficients describing changes in the hazard rate of time-to-retweet. The Cox proportional-hazards model estimates the survival rate of an object of study; in this case, how long before a tweet is retweeted. The less a tweet survives before being retweeted (i.e., the faster it is retweeted), the more the user accepts the content. In terms of the results given by the model, positive coefficients indicate an increase in the hazard rate (faster time-to-retweet) and acceptance of the frames. In contrast, negative coefficients indicate slower times and greater rejection of the frames.

One important aspect is that verified users neither posted nor retweeted racist messages.[23] This behavior is probably a direct result of the costs public figures incur when engaging with racist content, something that less-public users can get away with. Also note that many verified users are news outlets and government officials who often have staff managing their social media accounts and more to lose if they were to engage publicly in overt forms of racism.

Table 1 shows the results from the Cox proportional-hazards model. For an intuitive interpretation of the magnitude of the effect, consider the effect of our covariate of interest, *racist attack*, among users of the progovernment community, which is negative and takes the value of –0.148 ($p < 0.01$) in model 1. The exponentiated value of the coefficient (0.86) is the incidence rate and can be interpreted as the (instantaneous) change in *time-to-retweet* when a tweet has racist content. One minus the incidence rate is 0.14, showing that the *time-to-retweet* for racist posts is about 14 percent slower. Thus, users are not eager to retweet racist tweets posted by their own authorities, suggesting that the content of racist messages produces rejection among users, or that sharing racist messages increases the latency of the decision to bear the (social) cost of sharing racist message content. The survival probability curves of *time-to-retweet* for racist and nonracist

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

Table 1. Racist Content and Time-to-Retweet in the Ecuadorian Progovernment Community

	Model 1
Racist attack	–0.148***
	(0.012)
Auth bot control	–0.022***
	(0.008)
Hub bot control	–0.292***
	(0.028)
High degree auth (dummy)	–0.216***
	(0.004)
High degree hub (dummy)	0.085***
	(0.003)
Friends auth (ln)	0.149***
	(0.008)
Followers auth (ln)	0.095***
	(0.002)
Friends hub (ln)	0.241***
	(0.020)
Followers hub (ln)	0.065***
	(0.015)
N	465,511

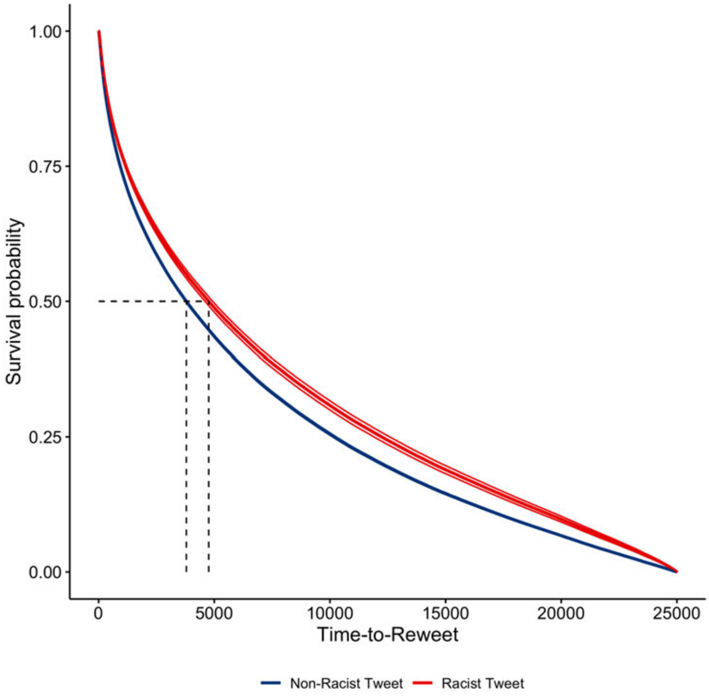
Standard errors are reported in parentheses, with confidence levels reported as follows: *p < .1; **p < .05; ***p < .01.
Note: Hazard estimates of time-to-retweet. Positive numbers indicate positive increases in hazard rate and shorter time to retweet.

content are plotted in figure 4. The results are in line with the expectations from hypothesis 1.

There are other particularities about the network worth mentioning. The progovernment community was not eager to retweet messages from its authorities. Results show that users more central to the network (authorities with an in-degree above the log median) were retweeted 20 percent more slowly than users less central to the network (authorities with an in-degree below the log median). However, users more central to the network retweeted posts 9 percent faster than users less central to the network. The difference in behavior shows that the "soldiers" of the network—i.e., users less central to the network—were not being

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

Figure 4. Survival Probability of Cox Proportional-hazards Modes



Note: From table 1, model 1, with time-to-retweet as the dependent variable. Survival curves with 95 percent confidence intervals.

activated by the content posted by the central users but rather by content created by other, less prominent users.

As anticipated by hypothesis 1, racist content had a higher time-to-retweet than nonracist content. The larger latency in *time-to-retweet* suggests that, in general, racist content induced rejection from users and a slower diffusion of the frame. Furthermore, the results suggest that public figures avoid producing racist content and avoid reproducing it. The social cost of overt racism is a likely deterrent for that behavior, especially damaging to the image of more prominent political personalities. Alternatively, more-public figures might also be less racist. While we cannot discard this possibility, it seems less plausible, considering the racist structure embedded in this network.

Threats to the In-Group and Racist Content

The main argument of this article is that racism is activated when the status of the in-group is threatened (H2). To provide empirical evidence, we can look at two events in

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

which the indígena community (out-group) threatened the status of the government, either by calling on forces to challenge the government or by calling the government to acquiesce to its demands. Specifically, I test the hypothesis by examining the moments when Jaime Vargas, the president of the CONAIE, called the police and military forces to disobey government orders and join the protest, and when the Moreno government announced the end of the strike after agreeing to the demands of the indígena leadership.

To determine whether these two events affected users' time-to-retweet to racist content, I followed Calvo et al. (2020) and employed an interrupted time series analysis, a variety of regression discontinuity design (RDD) in which the running variable is time (Morgan and Winship 2015; Mummolo 2018). The key assumption of this approach is that any immediate change in the time-to-retweet can be attributed to the events and not to any other factor affecting time-to-retweet that also changed systematically at the same point in time.[24] The granularity and abundance of Twitter data, as well as the public nature of these events, allows me to isolate the change in time-to-retweet on the specific second the event took place and thus make a more plausible assumption that other omitted variables (e.g., other events not attacking the status of the in-group) are not also changing suddenly at the time of the events. To further isolate the effect of the event, I estimated the models within a six-hour window around the cutoff (i.e., time of the events).[25]

As noted, the primary quantity of interest is the immediate change in time-to-retweet after the cutoff. I estimated the models using a nonparametric local linear regression (LLR) to approximate the treatment effect at the cutoff points (Gelman and Imbens 2019). In other words, I fitted two separate regression functions, one before and one after the cutoff events. The treatment effect (i.e., the effect of the event) is the difference of both functions at the limits of the cutoff.[26]

A possible threat to the model's validity is the potential of sorting by users right before the event. In this case, users might anticipate the outcome of the events (i.e., a perceived threat to the status quo) and change their behavior accordingly. Given that we expect the event to decrease *time-to-retweet*, any anticipation to the treatment is likely to go in the same direction. Therefore, the model would underestimate the effects of the treatment, suggesting that the true effect of the event would be stronger. However, I ran a McCrary test and found no evidence of sorting around the cutoff for racist users engaging with racist content (McCrary 2008). Appendix B provides additional tests to verify the continuity assumption, including placebo checks with the running variable. Overall, the results ensure the internal validity of the RD design. We now turn to the events and how they constitute a threat to the status of the progovernment group.

Calls to Arms and Acquiescing

Historically, the indígena uprisings have had three political outcomes: subjugation through violence, policy reform and expansion of rights (e.g., constitutional

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

recognition of Ecuador as a plurinational state), and coups. The 2019 protests began as a series of demands (i.e., policy reform) made by the indígena community in reaction to government action that culminated in eliminating gas subsidies. Indígenas began marching from their communities, often located in rural Ecuador or Amazonia, and headed, as is customary, to the capital, Quito, where they would occupy city spaces of symbolic importance and negotiate with the government.

On October 10, the indígena assembled in a large forum in Quito, where eight police officers, among them a colonel, were detained for more than two hours, as well as various journalists and their crews.[27] Jaime Vargas spoke to the crowd on live television (the detained broadcast journalists and their crews came in handy), with the detained police officers in the background. At 11:51 a.m., Vargas asked the detained colonel, the police force, and the military to join the protests and disobey the orders of the government.

The moment of Vargas's sedition created a reaction in the progovernment community. Given recent Ecuadorian political history, Vargas's call was not an empty threat but rather a clear challenge to the power and the status of the government.[28] Through Vargas's call, the out-group conspicuously challenged the in-group. We would expect this action to be perceived as a threat to the in-group and to trigger the progovernment community's shared identity, particularly the racial identity.

Similarly, on October 13, high-ranking officials from the Moreno government sat down with indígena leaders to negotiate the end of the strike. The indígena leaders demanded, among other things, that Moreno rescind the order that prompted the mobilizations in the first place. After three hours of negotiations, news stations broadcast the meeting live as each party made public statements. At 9:45 p.m., a United Nations mediator announced the conditions under which the strike would be terminated. Lenín Moreno agreed to reinstate the gas subsidies, and the indígena leaders agreed to suspend the mobilization. In the eyes of progovernment users, the indígenas had won. The adjudication of the strike materialized the challenges to the in-group status. As in the previous event, the immediate actions that led to the end of the strike were broadcast and garnered collective attention.

Furthermore, both episodes were highly salient events. High-salience events focus the public's attention and redefine the situation (Pride 1995. Lin et al. (2014, pg. 2) argue that highly salient media events "create a social condition called 'shared attention,' … where the larger potential audiences, altered norms, and the high level of *shared understanding* can all contribute to shifts in how users … attend to content" (emphasis added). Lin et al. (2014) explain that in media events, users can converse without needing to explain the context, allowing for the creation of a more disciplined message. In both of these events, progovernment users reacted to these perceived threats to the in-group status by engaging with racist frames.

The Effects of Perceived Threats on Racist Content

Table 2 reports the robust bias-corrected treatment effects and 95 percent confidence intervals as suggested by Calonico et al. (2014). The dependent variable is

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

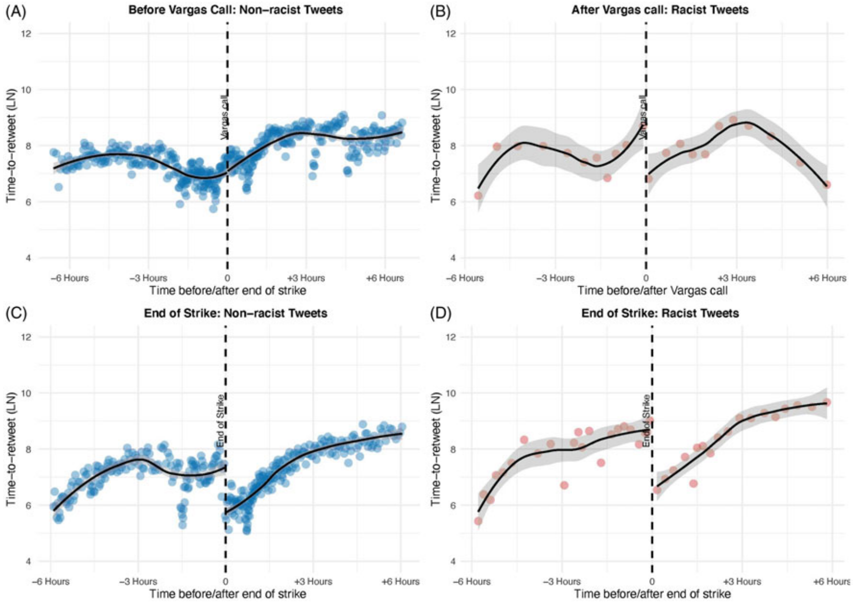
https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

Table 2. Regression Discontinuity Estimation for Racist and Nonracist Tweets

	Vargas's Call Model 2.A		End of Strike Model 2.B	
	Nonracist tweets	Racist tweets	Nonracist tweets	Racis tweets
Treatment effect	–0.030	–2.786**	–0.548**	–2.199**
	[–0.229, 0.169]	[–5.069, –0.502]	[–1.084, –0.012]	[–2.803, –1.435]
Number of observations	21,390	294	12,905	410
Pretreatment observations	11,183	174	4,864	143
Posttreatment observations.	10,207	120	7,941	267

Notes: 95 percent confidence intervals reported in brackets, with confidence levels reported as follows: *p < .1; **p < .05; ***p < .01.
Model 2.A centers treatment on October 10, 2019, at 11:51 a.m. local time, when Jaime Vargas, the president of the CONAIE, asked the police force and the military to join the protests and disobey the orders of the government. Model 2.B centers treatment on October 11, 2019, at 9:45 p.m. local time, when the United Nations mediator announced the terms agreed on for the end of the strike.

Figure 5. Time-to-Retweet During the Ecuadorian Protests



Top panels: 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. Bottom panels: Centering on October 11, 2019, at 9:45 PM, local time, when the United Nations mediator announced the terms agreed on for the end of the strike.

time-to-retweet for models 2.A and 2.B. In model 2.A, the cutoff time of the discontinuity is centered on October 10, 2019, at 11:51 a.m. local time, when Jaime Vargas, the president of CONAIE, asked the police force and the military to join the protests and disobey the orders of the government. In model 2.B, the time of discontinuity is centered on October 11, 2019, at 9:45 p.m. local time, when the United Nations mediator announced the terms agreed upon for the end of the strike. The treatment effect for racist tweets in both events is negative—time-to-retweet is reduced—and statistically significant at the 95 percent level.

To better illustrate the effect, panels a and b in figure 5 plot the regression discontinuity results for the event. For both panels, the vertical axis reports time-to-retweet, and it is interpreted as usual: lower values mean less time-to-retweet and less user latency. The sample is divided into racist and nonracist tweets. The horizontal axis has a range of 12 hours, 6 hours before and after the event. A LOESS smoother was used to fit the underlying regression function separately before and after the event. The discontinuity shows that there is no change in latency for users sharing nonracist content at the time of Vargas's speech. However, there is a statistically and substantively significant difference in latency

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

for users sharing racist messages ($p < 0.01$). Immediately after Vargas's call, racist content increases engagement and reduces latency among progovernment users. Notice that before the cutoff, racist content is spread, on average, slower than nonracist content. However, at the time of Vargas's speech, users engage with racist content faster than with nonracist content. The reaction to the perceived threat to the status of the in-group points to progovernment users' accepting racist frames and increasing the speed of their diffusion.[29]

Panels c and d in figure 5 show a similar treatment effect when the end of the strike was announced. In this case, the discontinuity shows that the end of the strike decreased the latency for both nonracist and racist content ($p < 0.05$), suggesting that adjudication (i.e., the act of granting or denying ownership of an outcome to groups in social media) also affects the reaction of users to content, a result in line with. Thus, some of the effects of the event on latency might be a product of adjudication. Nevertheless, the treatment effect for racist tweets (–2.17) is larger than the treatment effect for nonracist tweets (–0.66), suggesting that some of the change is a reaction to the threat to the in-group status.

Analogous to Vargas's call, racist content went from having a longer time to retweet than nonracist content to roughly the same time when the end of the strike was announced. In other words, both types of frames created similar acceptance from progovernment users after the strike adjudication.

Overall, I find that users of the in-group, the progovernment community, engaged with racist frames when the status of the in-group was perceived to be threatened. The analysis of two high-salience events shows reduced latency of engagement and greater acceptance from progovernment users of racist content. I also find that, absent these threats, racist content creates rejection from users.

Conclusions

This research has explored the spread of racist content on social media. Taking cues from the political psychology literature on out-group hate and the communications literature on online behavior, I expected racist frames to activate as a reaction to threats to the in-group. Given the social cost of racism and the real consequence of online activity on a user, there will be a variation in who reproduces racist messages and when these frames are activated.

This article contributes to the literature on various fronts. The analysis of the Ecuadorian protests of 2019 shows that perceived threats by the indígena community (out-group) to the status of the government (in-group) make progovernment users accept racist frames. While racist content was present in the study's Twitter network throughout the development of the protest, it required concrete threats, like the government's acquiescence to the indígena demands, for users to actively engage with racist frames. As Aruguete and Calvo (2018) show, shorter time-to-retweet is a proxy for cognitive congruence, suggesting that threats to the in-group status make racist frames congruent to in-group members.

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

Methodologically, this study provides an easy-to-implement process to identify racist tweets. It uses the machine-learning algorithm *Perspective* to detect identity attacks in our corpus. It combines these results with a list of terms that serve as markers for Ecuador's contextual forms of racism. Unlike alternative, semisupervised machine-learning approaches that require hard-to-come-by tagged corpora, *Perspective* is pretrained in various languages. By providing contextual information on the racist discursive forms, the program can effectively detect racism, avoiding many false positives. The method has some limitations: it cannot identify subtler forms of racism (which would be more reflective of the extent of racist ideology), and it overrepresents racism in communities where racism is unlikely to be produced. I believe that these are two areas where this process can improve in future research.

Polarized environments such as those found in Ecuador help detect the behavior of racist content, a problem that is pervasive in most political contexts. This study has identified some mechanisms that explain the proliferation of racist discourses in social media, which is essential to know so as to stop them eventually.

Aggressiveness and toxicity are not endemic to the community representing the dominant group in society. In this network, high levels of both can be found in the indígena community. However, aggressiveness and toxicity toward power are less problematic than the other way around. In the case of Ecuador, the racist ideology has created a complex relationship between the political elite and the indígena community, sometimes excluding them from participating in the political process (Hall and Patrinos 2004), other times forming (short-lived) political alliances in exchange for policy gains (Dávila Gordillo 2021). The network analysis reveals how, in this context, racism is propagated in social media, a mechanism that has yet to be systematically studied in an online environment. The research design's gains in internal validity also suggest further research to explore how generalizable the results are. Future research should focus on how racist content and racism in social media operate in other contexts and other outlets.

Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/lap.2022.25

Acknowledgement. This work was completed in part with resources provided by the Research Computing Data Core at the University of Houston. I received invaluable feedback from the members of the interdisciplinary Lab for Computational Social Science (iLCSS-UMD), as well as important suggestions from Ernesto Calvo, Francisco Cantú, Nikolay Marinov, SoRelle W. Gaynor, Diana Dávila Gordillo, and Joan C. Timoneda.

Notes

1. I also find evidence that threats to the status of the government community increase the diffusion of racist content (see appendix C).

2. In addition to identifying as individuals, people also identify as members of social groups to which they belong—that is, in-groups. People who identify as members of the "other" social groups are the out-group.

3. Many research frameworks examine more covert representations of racism, including laissez-faire racism (Bobo et al. 1997), color-blind racism, or no-difference racism (Bonilla-Silva 2003), and ventriloquism (Guerrero 1997). However, the discursive manifestations of these forms of racism are challenging to identify systematically.

4. A couple of days after the announcement, the government negotiated a deal with the Unión de Transportistas and ended the strike.

5. Similar to other instances of social mobilization (Aruguete and Calvo 2018; Bastos et al. 2015), media accounts of the 2019 protests in Ecuador (Roa Chejín 2019) suggest that Twitter was part of the political battleground. Note, however, that only 12.7 percent of Ecuadorians have a Twitter account compared to, for example, 67.3 percent who have a Facebook account (AmericasBarometer 2018). Furthermore, social media users in Ecuador are more likely to be young, urban, educated, and well off. For the data used in the paper, Twitter data were beneficial, despite the limitations on both sample and representation.

6. The analysis was restricted to retweets. Other forms of agreement expressed on Twitter, such as "likes," do not reproduce the original content in other users' feeds. The diffusion of frames is produced when content is shared (i.e., when tweets are retweeted).

7. The random-walk community detection algorithm is based on the premise that a cluster of nodes closely linked will create subgraphs in networks. Thse communities can be detected by estimating short, random walks between nodes. The short walks will remain in the same cluster.

8. I estimated a layout of node coordinates using the Fruchterman-Reingold (FR) force-directed algorithm in R 3.5 *igraph* (Csardi and Nepusz 2006). The FR algorithm facilitates the visual inspection of the network, communicating information about the proximity between nodes (data reduction pull) while preventing nodes from overlapping (force-directed push). For visualization, the data-reduction pull will cluster nodes from the same community, and force-directed push will avoid nodes' overlapping and reduce the number of overlapping edges.

9. Most of the external dialogue of indígena community users was with pro-Correa community users.

10. Google's API *Perspective* considers an identity attack any "negative or hateful comment targeting someone because of their identity." https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US.

11. 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*.

12. Lowering or raising the scores does not change the main outcomes, but it does change the accuracy of the model (see appendix A).

13. For a complete list of terms, see appendix A.

14. For a detailed recount of the process and comparative performance of the models, see appendix A.

15. F1, recall, and precision scores are performance metrics. Recall is the number of correct predictions divided by the total number of elements that should have been predicted. Precision is the number of correct predictions divided by the number of all returned predictions. F1 combines both scores and estimates the harmonic mean. High values for all three scores point to a more accurate model. A high precision score means few false positives in relation to the number of true positives.

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

16. For example, "He said that they are idiots. That is not true." is considered highly toxic by the *Perspective* algorithm. Appendix A explains how the *Perspective* algorithm works and the reason behind false positives.

17. In the indígena community, the algorithm detected 0.005 percent of racist tweets, compared to 2 percent from the progovernment community.

18. Twitter rules include the following: "You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease." A violation of this rule can result in the elimination of the tweet.

19. To reduce the chance of losing deleted tweets, I ran Twitter's backward search multiple times during the period analyzed.

20. For example, if we argue that users will take less time to retweet racist content, yet some racist tweets are misclassified as not being racist, these misclassified tweets will reduce the time difference between racist and nonracist content. Thus, it would be harder to reject the null hypothesis.

21. There is an interesting discussion of whether the total count of diffusion of a message or the response time to a message is the critical variable in the analysis (see Aruguete and Calvo 2018). I consider *time-to-retweet* to be closer to the cognitive intent to share. The count of times users diffuse a message depends on various elements, not only on agreement with the content: the number of people connected at the time of the tweet, how much of a specific frame is produced, and who is producing the content. Thus, testing the total count of diffusion is not only capturing agreement with the content but also other elements. However, I also tested whether threats to the in-group increased the probability of a retweeted post being racist. As expected, users retweeted more racist content (increased diffusion) when there was a threat to the in-group (see appendix C).

22. Bots in Twitter are accounts that autonomously create or reproduce tweets. They are often used to spread coordinated messages on the platform. They are created en masse and, to increase impact, follow thousands of users (friends). Bots themselves, however, have few followers. Therefore, to control for the effect of these accounts, I estimated the ratio of followers to friends for each user in the dataset.

23. Verified users in the Ecuadorian network are also high-degree users. This is to be expected.

24. More formally, we must assume continuity in the potential outcome functions at the treatment boundary (Morgan and Winship 2015).

25. This is the same window used by Calvo et al. (2020), who use Twitter data to determine the effect of adjudication. Mummolo (2018), who employs an interrupted time-series analysis to estimate how procedural changes in police affected the rate of stops of criminal suspects, uses a 120-day window.

26. To determine the bandwidth size of the RDD, I followed Calonico et al. (2014). They use a data-driven mean-squared error search to select optimal bandwidths and a triangular kernel that assigns linear downweighting to each observation. Appendix B tests different bandwidths to show that the effect is not driven by different bandwidth sizes.

27. The details of the detainment are unclear. The indígena community blamed the police and the government for the deaths of various indígena protesters. This was the main reason for holding the police officers. The government said the police officers were kidnapped, while the indígena leaders, mainly Jaime Vargas, argued that the officers and the press could leave at any time.

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

[28] For example, in the 2000 coup that unseated President Jamil Mahuad, then–CONAIE president Antonio Vargas was joined by the highest-ranking army general and the president of the Supreme Court in a triumvirate that briefly held power.

[29] Appendix C shows evidence suggesting that, in addition to the speed of diffusion of racist content, Vargas's call also increased the probability of users' retweeting racist content.

References

AmericasBarometer. 2018. Dataset. *The AmericasBarometer by the LAPOP Lab*. Vanderbilt University: LAPOP. www.vanderbilt.edu/lapop

Amira, Karyn, Jennifer Cole Wright, and Daniela Goya-Tocchetto. 2021. In-Group Love Versus Out-Group Hate: Which Is More Important to Partisans and When? *Political Behavior*, 43, 2: 473–494.

Aruguete, Natalia, and Ernesto Calvo. 2018. Time to #Protest: Selective Exposure, Cascading Activation, and Framing in Social Media. *Journal of Communication* 68, 3: 480–502.

Barberá, Pablo, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. Tweeting from Left to Right: Is Online Political Communication More than an Echo Chamber? *Psychological Science* 26, 10: 1531–42.

Barlett, Christopher P., Caroline C. DeWitt, Brittany Maronna, and Kaleigh Johnson. 2018. Social Media Use as a Tool to Facilitate or Reduce Cyberbullying Perpetration: A Review Focusing on Anonymous and Nonanonymous Social Media Platforms. *Violence and Gender 5*, 3: 147–52.

Bastos, Marcos, Dan Mercea, and Arthur Charpentier. 2015. Tents, Tweets, and Events: The Interplay Between Ongoing Protests and Social Media. *Journal of Communication* 65, 2: 320–350. https://doi.org/10.1111/jcom.12145

Beck, Scott H., Kenneth J. Mijeski, and Meagan M. Stark. 2011. ¿Qué es racismo? Awareness of Racism and Discrimination in Ecuador. *Latin American Research Review* 46, 1: 102–125.

Becker, Marc. 2010. *Pachakutik: Indigenous Movements and Electoral Politics in Ecuador*. Lanham: Rowman & Littlefields.

Bobo, Lawrence, James R. Kluegel, and Ryan A. Smith. 1997. Laissez-faire Racism: The Crystallization of a Kinder, Gentler, Antiblack Ideology. *Racial Attitudes in the 1990s: Continuity and Change* 15: 23–25.

Bonilla-Silva, Eduardo. 2003. Racial attitudes or racial ideology? An alternative paradigm for examining actors' racial views. *Journal of Political Ideologies* 8, 1: 63–82. https://doi.org/10.1080/13569310306082

——. 2015. The Structure of Racism in Color-blind, "Post-racial" America. *American Behavioral Scientist* 59, 11: 1358–1376.

Bretón, Víctor, and Francisco García Pascual. 2003. *Estado, etnicidad y movimientos sociales en América Latina: Ecuador en crisis*. Barcelona: Icaria.

Brewer, Marilynn B. 1999. The Psychology of Prejudice: Ingroup Love and Outgroup Hate? *Journal of Social Issues* 55, 3: 429–44.

Brock, Andre. 2009. Life on the Wire: Deconstructing Race on the Internet. *Information, Communication & Society* 12, 3: 344–63.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica* 82, 6: 2295–2326.

Calvo, Ernesto, and Natalia Aruguete. 2020. *Fake news, trolls y otros encantos: cómo funcionan, para bien y para mal, las redes sociales*. Madrid: Siglo XXI.

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

Calvo, Ernesto, Silvio Waisbord, Tiago Ventura, and Natalia Aruguete. 2020. "Winning! Electoral Adjudication and Dialogue in Social Media." Working paper.

Chatzakou, Despoina, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali 2017. Mean Birds: Detecting Aggression and Bullying on Twitter. In *Proceedings of the 2017 ACM on Web Science Conference (WebSci '17).* Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/3091478.3091487

Csardi, Gabor, Tamas Nepusz, et al. 2006. The Igraph Software Package for Complex Network Research. *InterJournal, Complex Systems* 1695, 5: 1–9.

Daniels, Jessie. 2013. Race and Racism in Internet Studies: A Review and Critique. *New Media & Society* 15, 5: 695–719.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *arXiv preprint. arXiv*: 1703.04009.

Dávila Gordillo, Diana L. 2021. Surviving Against All Odds: Pachakutik's Electoral Support, Mobilization Strategies, and Goal Achievement Between 1996 and 2019. Ph.D. diss., Leiden University.

De la Torre Espinosa, Carlos. 1996. El racismo en el Ecuador: experiencia de los indios de clase media. In *El racismo en el Ecuador: Experiencia de los indios de clase media*. Buenos Aires: CLACSO.

Díaz, Isabel, and Adriana Mejía Artieda. 2020. Las elites en octubre: de ciudadanos indignados a propietarios alarmados. In *Octubre y derecho a la resistencia*, ed. Franklin Ramírez Gallegos. Buenos Aires: CLACSO. 271–86.

ElSherief, Mai, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to Peer Hate: Hate Speech Instigators and Their Targets. *arXiv preprint arXiv:1804.04649.*

Eschmann, Rob. 2020. Unmasking Racism: Students of Color and Expressions of Racism in Online Spaces. *Social Problems* 67, 3: 418–36.

Gelman, Andrew, and Guido Imbens. 2019. Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs. *Journal of Business & Economic Statistics* 37, 3: 447–46.

Groenendyk, Eric W., and Antoine J. Banks. 2014. Emotional Rescue: How Affect Helps Partisans Overcome Collective Action Problems. *Political Psychology* 35, 3: 359–78.

Guerrero, Andrés. 1997. The Construction of a Ventriloquist's Image: Liberal Discourse And The "Miserable Indian Race" in Late 19th-century Ecuador. *Journal of Latin American Studies* 29, 3: 555–90.

Hall, Gillette, and Harry Anthony Patrinos. 2004. *Indigenous Peoples, Poverty, and Human Development in Latin America, 1994–2004*. Washington, DC: World Bank.

Himelboim, Itai, Marc Smith, and Ben Shneiderman. 2013. Tweeting Apart: Applying Network Analysis to Detect Selective Exposure Clusters in Twitter. *Communication Methods and Measures* 7, 3–4: 195–223.

Hosseini, Hossein, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint. arXiv:1702.08138.*

Huddy, Leonie. 2001. From Social to Political Identity: A Critical Examination of Social Identity Theory. *Political Psychology* 22, 1: 127–56.

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press

Huddy, Leonie, Lilliana Mason, and Lene Aarøe. 2015. Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity. *American Political Science Review* 109, 1: 1–17.

Jaidka, K., Zhou, A., Lelkes, Y., Egelhofer, J. and Lecheler, S. 2022. Beyond Anonymity: Network Affordances, Under Deindividuation, Improve Social Media Discussion Quality. *Journal of Computer-Mediated Communication* 27, 1: p.zmab019.

Kearney, Michael. 2019. rtweet: Collecting and analyzing Twitter data. *Journal of Open Source Software* 42, 4: 1829. https://doi.org/10.21105/joss.01829, R package version 0.7.0, https://joss.theoj.org/papers/10.21105/joss.01829

Keipi, Teo, Matti Näsi, Atte Oksanen, and Pekka Räsänen. 2017. *Online Hate and Harmful Content: Cross-national Perspectives*. London: Taylor & Francis.

Lapidot-Lefler, Noam and Azy Barak. 2012. Effects of Anonymity, Invisibility, and Lack of Eye-Contact on Toxic Online Disinhibition. *Computers in Human Behavior* 28, 2: 434–43.

Lin, Yu-Ru, Brian Keegan, Drew Margolin, and David Lazer. 2014. Rising Tides or Rising Stars? Dynamics of Shared Attention on Twitter During Media Events. *PLoS ONE* 9, 5: e94093. https://doi.org/10.1371/journal.pone.0094093

Martínez-Echazábal, Lourdes. 1998. *Mestizaje* and the Discourse of National/Cultural Identity in Latin America, 1845–1959. *Latin American Perspectives* 25, 3: 21–42.

Mason, Lilliana. 2016. A Cross-cutting Calm: How Social Sorting Drives Affective Polarization. *Public Opinion Quarterly* 80, S1: 351–77.

McCrary, Justin. 2008. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics* 142, 2: 698–714.

Morgan, Stephen L., and Christopher Winship. 2015. *Counterfactuals and Causal Inference*. Cambridge: Cambridge University Press.

Mummolo, Jonathan. 2018. Modern Police Tactics, Police-Citizen Interactions, and the Prospects for Reform. *Journal of Politics* 80, 1: 1–15.

Nakamura, Lisa. 2007. *Digitizing Race: Visual Cultures of the Internet*. Minneapolis: University of Minnesota Press.

Neumayer, Christina, Luca Rossi, and Björn Karlsson. 2016. Contested Hashtags: Frankfurt in Social Media. *International Journal of Communication* 10: 22 pp.

Pons, Pascal, and Mattheiu Latapy. 2005. Computing Communities in Large Networks Using Random Walks. In Computer and Information Sciences - ISCIS 2005. ISCIS 2005. Lecture Notes in Computer Science, 3733. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11569596_31

Preussler, Annabell, and Michael Kerres. 2010. Managing reputation by generating followers on Twitter. In Medien–Wissen–Bildung Explorationen visualisierter und kollaborativer Wissensräume: 129–143.

Pride, Richard A. 1995. How Activists and Media Frame Social Problems: Critical Events Versus Performance Trends for Schools. *Political Communication* 12, 1: 5–26.

Roa Chejín, Susana. 2019. La pelea por la apariencias tuiteras. GK. October 23. https://gk.city/2019/10/21/hashtags-paro-nacional-ecuador/

Roitman, Karem. 2009. *Race, Ethnicity, and Power in Ecuador: The Manipulation of Mestizaje*. Boulder: FirstForumPress.

Roitman, Karem, and Alexis Oviedo. 2017. Mestizo Racism in Ecuador. *Ethnic and Racial Studies* 40, 15: 2768–86.

Schmidt, Anna, and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on*

Natural Language Processing for Social Media, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Sue, Christina A., and Tanya Golash-Boza. 2013. "It Was Only a Joke": How Racial Humour Fuels Colour-Blind Ideologies in Mexico and Peru. *Ethnic and Racial Studies* 36, 10: 1582–98.

Tajfel, Henri. 1981. *Human Groups and Social Categories: Studies in Social Psychology.* Cambridge: Cambridge University Press.

Timoneda, Joan C. 2018. Where in the World Is My Tweet: Detecting Irregular Removal Patterns on Twitter. *PloS one* 13, 9: e0203104.

Van Cott, Donna Lee. 2008. *Radical Democracy in the Andes*. Cambridge: Cambridge University Press.

Van Dijk, Teun Adrianus. 2005. *Racism and Discourse in Spain and Latin America*. Amsterdam: John Benjamins.

Wulczyn, Ellery, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *WWW '17: Proceedings of the 26th International Conference on World Wide Web*. Geneva: International World Wide Web Conferences Steering Committee. 1391–99.

Supporting Information

Additional supporting materials may be found with the online version of this article at the publisher's website: Appendix.

https://doi.org/10.1017/lap.2022.25 Published online by Cambridge University Press