Semi-Supervised Classification of Overt and Covert Racism in Text

Abstract

Extant work has identified two discursive forms of racism: overt and covert. While both forms have received attention in scholarly work, research on covert racism has been limited. Its subtle and context-specific nature has made it difficult to systematically identify covert racism in text, especially in large corpora. In this article, we first propose a theoretically driven and generalizable process to identify and classify covert and overt racism in text. This process allows researchers to construct coding schemes and build labeled datasets. We use the resulting dataset to train XLM-RoBERTa, a cross-lingual model for supervised classification with a cutting-edge contextual understanding of text. We show that XLM-R and XLM-R-Racismo, our pretrained model, outperform other state-of-the-art approaches in classifying racism in large corpora. We illustrate our approach using a corpus of tweets relating to the Ecuadorian *indígena* community between 2018 and 2021.

Warning: Due to the nature of the topic, this paper contains offensive content.

1 Introduction

Scholars theorizing about the different discursive forms taken by racism have identified two clear patterns: overt and covert racism (Van Dijk, 2005; Bonilla-Silva, 2006; Coates, 2011). While overt racism refers to direct attacks using language openly acknowledged as derogatory (e.g., slurs), covert racism can be described as "subtle, and subversive institutional or societal practices, policies, and norms utilized to mask the structural racial apparatus" (Coates, 2008, p.212). As its name suggests, covert racism has a hidden nature that has brought scholars to focus on the social practices that embody it rather than on its discursive forms (Chin, 2015; Coates, 2008, 2011; Hampton, 2010; Holdaway and O'neill, 2007; Levchak, 2018; Marom, 2019; Walsh, 2004).

Researchers have emphasized the performative nature of covert racism, highlighting the importance of context to identify these practices. Covert racism is a style of discrimination that is hard to detect and even harder to label. It entails the use of language, rhetorical strategies and "semantic moves" aimed at avoiding explicit racial conversations and racist language (Essed, 1991; Bobo, Kluegel and Smith, 1997; Bonilla-Silva, 2015). Thus, to understand these practices as racist, as well as their discursive manifestations, we need an understanding of the historical and sociological processes underpinning the racist ideology embedded into the social structures of the place of study. Given the complexity of identifying covert racism in general, most scholars focus on racist practices rather than on racist *discourse*. Extant research on covert racism addresses mainly anecdotal observations of these practices (Holdaway and O'neill, 2007; Levchak, 2018) or interviews (Bonilla-Silva, 2002; Marom, 2019; Shoshana, 2016), rather than on other sources of abundant information, such as large corpora.

In this paper we address two gaps in this literature. First is the paucity of research on covert racism in *discourse*. Second is the limitations found in current computational tools to systematically identify overt and covert racist content in text, especially in large corpora. To address the first gap, we propose a generalizable step-by-step process to identify and classify covert and overt racist discourse. The process explicitly integrates contextual knowledge to the development of lexical rules to categorize speech acts.

Taking the Ecuadorian case as an example, we first identify covert and overt racist practices based on existing literature. We then transform the narrative of these practices into lexical rules to classify discourse. Covert racist discourse is pervasive but requires an understanding of the context in which it is produced and reproduced. In particular, it is necessary to include an understanding of the historical constructions of hierarchies produced by racist ideologies and the forms of domination that are manifested in social settings (including language). Constructing rules to identify racist discourse using this context allowed us to detect in the text the (covertly) racist practices described by scholars. We use these rules to manually label a dataset to train a supervised machine learning model.

Second, work on hate speech detection has highlighted how models trained to detect hate speech, including implicit hate speech, fail to generalize to other corpora (Kim, Park and Han, 2022; Kim et al., 2023), as well as the lack of clear definitions for concepts like 'implicit and explicit hate speech' (Ocampo et al., 2023). We address this limitation from the computational approaches, not by proposing a generalizable model, but rather by proposing a generalizable approach, that acknowledges the context-dependent nature of racist language, and how to take this into account when constructing the training sets and further pre-training Transformers-based models. To this end, we apply a Natural Language Processing (NLP) technique for contextual understanding of text to classify covert and overt racism. While conventional NLP techniques are unable to fully understand contextual cues from text,[1] Transformer-based models (e.g., BERT) have greatly improved contextual understanding of text (see Timoneda and Vallejo Vera, 2025*a*). In this article, we further pretrain and fine-tune a Transformers-based cross-lingual RoBERTa model, XLM-R, to detect overt and covert racism in Spanish. We place special emphasis on building a theoretically-grounded training set and further pre-training the XLM-R model with 1.7 million tweets that include racist language specific to the Ecuadorian context. We then compare XLM-R performance with three sets of widely used machine learning and deep learning approaches. First are traditional

[1] Examples of these approaches are SVMs or Logistic Regression in machine learning or Convolutional Neural Networks (CNNs) and Long Short Term Memory Networks (LSTMs) combined with word embeddings in deep learning. LSTMs integrate some contextual awareness by allowing less immediate words in the text to remain important, but they still underperform in cases where the classifier demands subtle, contextual information such as detecting covert racism.

machine learning models such as Support Vector Machines (SVM) and logistic regression. Second are two neural networks, one convolutional (CNN) and one recurrent (Bi-LSTM), which we pair with two sets of word embeddings. Third, we test the performance of two generative Large Language Models (LLMs), OpenAI's GPT-4o and Meta's Llama-3.3-70B, using few-shot learning. We find that our models significantly outperform these alternative approaches in our annotation task.

The remainder of the paper is organized as follows. In section 2 we give an overview of the literature focusing on overt and covert racism. We concentrate on research that studies the discursive manifestation of overt and cover racism. In the following section, we introduce our step-by-step process and detail the steps taken to define the coding rules for the training data (using Ecuador as an example). In section 4 we discuss the limitations of non-contextual machine learning approaches to classifying racist content, and how these can be solved by context-rich models such as RoBERTa. We then apply our lexical rules to label a training set (of tweets), and train a RoBERTa-based model to categorize a large corpus (+2M tweets). In section 6 we show the prevalence and activation of covert and overt racism in the Ecuadorian Twitter network during the 2019 *indígena* protests. We conclude this research by addressing the limitations of our approach and possible extensions.

2 Covert and Overt Racism

Although we are particularly interested in discursive manifestations of racism, doing so requires to first engage, though briefly, with a general understanding of racism writ large, as well as non-discursive manifestations of racism. Scholars have conceptualized and understood racism differently, often as a complex and evolving notion. Common to many authors is the idea of racism as a system of racial domination sustained by racial domination projects (Bonilla-Silva, 2001; Van Dijk, 2009). These projects create *races* out of people who were not so before: European peoples becoming White (Painter, 2010), African peoples becoming Black (Wright, 2004), and peoples of the Americas becoming *Indios* (Bonfil Batalla, 1977). Racism is a combination of practices and behaviors that produce and reproduce racial structures and define social relations among newly created races leading to economic, political, and, overall, material consequences (Bonilla-Silva, 2015).

Racism is manifested in different forms: marginalization, discrimination, exclusion, and violence, among other social practices that go beyond discourse. These practices are the product of domination–power of one group over another–, and result in distinct social inequalities (Van Dijk, 2009). The structural and systemic dimension of racism means that its manifestations are at the individual (Kinder and Sears, 1981; Henry and Sears, 2002) and institutional level (Bridges, 2019). For the former, through discriminatory practices and individual expression; for the latter, through institutions–e.g., public policy, laws, and access to political spaces (Ahmed, 2012). Broadly, racist discourse reflects how people communicate these practices and how they view "power relations, normative frameworks, and the combination of symbolic and material reality" (Herzog and Lance Porfillio, 2022); racist discourse is directly involved in the reproduction of the racist social structure (see Van Dijk, 1993, 2005).

Covert racism operates through a process of socialization that creates distance between the dominant and the dominated group in a social structure. Covert racism creates boundaries that are assumed "natural, legitimate, and normal" (Coates, 2008, p.111), while perpetuating and reinforcing the racial order. It embodies a series of social practices, at the individual level–e.g., shunning (Xu, 2025)– and group level–e.g., redlining (Lynch et al., 2021)–, informal–e.g., racial profiling (Glover, 2009)– and formal–e.g. benign neglect (Elias, 2024)–.

As racist practices can be both overt–e.g., Jim Crow in the United States or Apartheid in South Africa–and covert–e.g., racialized minorities targeted by police officers (Pierson et al., 2020)–, so can racist discourse. Overt racism consists of direct attacks using language openly acknowledged as derogatory (e.g., slurs), taking an "open and blatant nature" (Coates, 2011). Covert racist language, on the other hand, is "hidden, secret, private, covered, disguised, insidious, or concealed" (Coates, 2011), understood as a set of linguistic practices that advance racialized differentiations and stratification that rely on unspoken assumption and expectations about the racist structure (Kroskrity, 2021).

Studying the discursive forms of covert racism, as well as with any form of racism, requires deep knowledge of the context in which this discourse is employed. The discursive manifestations of covert racism are largely *context dependent.* Racial stereotypes change from context to context as there are "differential expectations for different racial groups" (Coates, 2008, p.213). The importance of context lies in the fact that covert racism perpetuates a particular racist social order.[2] Its discursive form thus reflects this specific social order, which may need to be more evident to untrained (de-contextualized) eyes.[3]

Still, within the contextual nature of covert racism, scholars have identified certain generalizable characteristics of this discourse. First, it reproduces racial stereotypes with slightly sanitized terms (Shoshana, 2016). For example, the representations of Asian-Americans as "nerds" within the United States. The terms are innocent enough to make racial elites able to participate in such practices (Coates, 2008, p.222). Second, covert racism relies on racial codes, that is, "words, phrases and/or ideas that may camouflage its true racist intent or purpose" (Chin, 2015; Coates, 2011; Hill, 1995). These codes are often presented (and represented) by local authority figures that use the language to perpetuate the practices of domination (Bonilla-Silva, 2002; Coates, 2011; Levchak, 2018). Third, covert racism is presented under a political correctness and "politeness" façade (Bonilla-Silva, 2002; Coates, 2011). As Bonilla-Silva (2002) argues, this form of covert racism (defined by the author as "color-blind racism") is presented in a slippery and ambivalent form. In some cases, speakers preface racist statements with different forms of *disclosure of non-racism* ("I am not a racist...", "I have [racial identity] friends...") or clarify that the topic pertains to anything but race. Alternatively, speakers take both sides (the non-racist and racist sides), or aim to soften racist statements with diminutives (Bonilla-Silva, 2002). The discursive ambivalence, and the semantic moves employed, can take different shapes under different

[2] In different spaces, covert racism can operate as reactions to historical developments: in the U.S. from affirmative action (Kinder and Sanders, 1996), in Europe from immigration (Blinder, Ford and Ivarsflaten, 2013), and in Brazil from the myths surrounding their 'racial democracy' (Telles and Bailey, 2013). When covert racism is the product of changes in the racial structure, then the development of a racist structure and its manifestations are contextual to the place and time analyzed. For instance, the colonial consequences of the Spanish legacy in Latin America developed a racial/racist structure different from that imposed in British colonies (Katzew, 2005; Rich et al., 1990).

[3] Of course, those who do not want to see such racist forms may also benefit from selective blindness.

contexts.

Fourth, covert racism discourse often takes the form of light talk –a hybrid between private sphere and public sphere statements (Hill, 1995). These discursive forms presented as light talk hide racist content behind a veneer of playfulness (Levchak, 2018). For example, the use of Spanish words in English language statements with ambivalent meanings, e.g. "*adios*" or "*hasta la vista*" which could mean a formal farewell or a form of expulsion (Hill, 1995). Finally, covert racism can take the form of "microinvalidations". In this case, statements negate experiences or negate race entirely. Individuals are told not to be "too sensitive" or the speaker declares herself as "not able to see race" (Bonilla-Silva, 2002; Levchak, 2018; Shoshana, 2016; Walsh, 2004).

It is important to note that this is not an exhaustive list of generalizable characteristics, but rather some regularly mentioned across cases. Beyond the five enumerated above, there are many characteristics of covert linguistic racism that are common in some contexts but not in others. For example, ventriloquism (which we explain in detail below), is a covert trait found in some countries in Central Africa (Murrey and Jackson, 2020) and the Andes (Martínez Novo, 2018), but not mentioned in other cases. Additionally, there are other types of behavior that can be used to hide the meaning of a message, among them racist messages, though not exclusively. Dog whistling, for example, is the use of coded language with the objective to signal group membership or support for an issue without having to do it openly (Albertson, 2015). As covert racism, the use of dog whistling allows the speaker to present their ideas at a lower risk. Yet, while it can be associated with racist intent, it can also be used to signal support for conspiracy theories which do not include racial content (Blackington and Cayton, 2025), signal socio-economic discrimination (Kruk et al., 2024*a*), or use religious appeals (Albertson, 2015).[4] As we stress throughout the article, contextual knowledge is paramount when developing an understanding of the (covertly) racist discourse of the case(s) analyzed.

[4] Dog whistles are about the double meaning–sending the right message to the right people while also keeping a façade for those in the out-group. They are, in effect, used to avoid the costs (social and otherwise) of making some statements public (Blackington and Cayton, 2025). However, these do not necessarily relate to racist behavior. Covert racism, by contrast, need not include the intention of message-rearing, yet it engaging with racist ideas and positions.

2.1 Covert and Overt Racism in Ecuador

Ecuador serves as an interesting example of a racial structure that emerged from mestizaje and Spanish colonialism, similar to the rest of Latin America. Through this case, we exemplify how to contextualize racism to a place and time, and how we can apply our categories of interest (e.g., covert and overt racism), to this context. In their own work, researchers need to determine their target context of study–place and time–, and adapt this and the next steps accordingly.

The racial structure in Ecuador is a product of the Spanish colonial domination project. The material and political inequalities began with the *hacienda* or *huasipungo* system in the late 16th century, where Indians were expected to pay tribute to the landowners in exchange for the right to use a parcel of land in the hacienda's territory (Oberem, 1985). After independence in 1822, the *hacienda* system, as well as the *indio* tribute, continued as large landholdings expanded. Only during the 1960s, with the land reform, oil boom, early industrialization, and expansion of the urban center did the *hacienda* system wither out (Pallares, 2002). Nevertheless, the racial structure remained. Thus, the indígena population started to organize and mobilize politically. This organization took the shape of the indígena uprisings of the nineties, massive mobilizations throughout the 21st century and the foundation and institutionalization of a political party. The indigenous population demanded and eventually conquered important political and social victories (see Pallares, 2002).[5] Yet, rejection of their demands from the mestizo and blanco-mestizo population has been constant, as have racist attacks and material inequalities.[6]

The state structures in Ecuador have perpetuated unequal systems dominated by a blanco-mestizo dominant ideology (Roitman and Oviedo, 2017). Research on race and ethnicity in Ecuador shows that *mestizaje* has created a 'whitening' process. Mestizaje, and also most people who identify as mestizos, downplays discrimination towards indígenas and other marginal communities while still engaging in *covert* racist practices (Beck, Mijeski and Stark, 2011). Unsurprisingly, covertly racist language is

[5] The indígena population has organized around the "indígena" identity. Note, however, that different and sometimes conflicting groups form it.

[6] We provide a more nuanced historical look at the racial structure in Ecuador in Appendix A.

normalized while having different forms. Individuals and society at large cover the racist elements of their actions and discourses through rhetorical means (Traverso-Yépez, 2005), which leads to variation in the manifestations of racist discourse.

Racism in Ecuador, deeply rooted in European colonialism, manifests as a "system of ethnic-racial dominance" (Van Dijk, 2009), primarily impacting the indigenous population. Despite their political activism and historical gains, indigenous communities continue facing marginalization due to a state that offers minimal support and unequal political access (De la Torre, 1996). In Ecuador, all institutions, classes, and contexts perpetuate a subtle yet pervasive blanco-mestizo racist ideology, often dismissed by its users as non-racist (Roitman and Oviedo, 2017). This makes covert racism pervasive, which opens an ideal opportunity to test our method.

2.2 Discursive Manifestations of Racism in Ecuador

We now turn to the task of identifying covert and overt forms of racist discourse considering its historical, social, and political context in Ecuador. To do this, we surveyed the extant literature on racist practices and discursive forms specific to our case. We started by asking how racial structure manifests itself in modern-day Ecuador. As discussed above, researchers have recently focused on the indigenous population's organization. Hence, this article builds on prior research on the social and political standing of the indigenous population in Ecuador. These works highlight that a) diverse forms of racism, racial stereotyping, and racial discrimination are entrenched in Ecuadorian society (Beck, Mijeski and Stark, 2011), and b) that the intersection of ethnicity, race, and class has created spaces where race is used both as a mobilizer and as a social marker (Whitten Jr, 2003). Additionally, we make use of more recent work focusing solely on racist practices and discourse to ensure we cover the breadth and depth of research on the topic. Our work was to systematize and bring together their insights, applying our definitions of covert and overt racism.

For our case, we define overtly racist language as any discourse that falls within one of the following categories: 1) discourse that explicitly includes derogatory terms or phrases that have been historically used to characterize the indígena population (either as individuals or as a community) as the lesser and dominated group in Ecuadorian society; 2) insults directed towards members of the indígena community that explicitly include their identity; 3) aggressive or denigratory language that includes the word 'indio'; 4) racialized phrases or idioms; and 5) violence or incitement to violence towards members of the indígena community. Additionally, we define covert racist language as any discourse that describes the actions or character of the indígena population (either as individuals or as a community) by reproducing the idea of them as the lesser and dominated group in Ecuadorian society through masked, sanitized, or de-racialized language. The codebook used is available in Appendix C, where we provide details and examples of each category.

Some sources referenced *did not* use our terminology (i.e., covert and overt racism) to describe the manifestations of racism. Nonetheless, we parsed the manifestations of racism studied by these scholars following our definitions of covert and overt racism. Overall, we identify five general manifestations of covert racism: 1) *no-difference racism* (Bonilla-Silva, 2006) or negation of identity (Canessa, 2007); 2) *attacks on the capabilities of the indigenous population* (Roitman and Oviedo, 2017); 3) *infantilization* of the indígena population; 4) *hygienic racism* or deeming the indígena population unclean (metaphorically or literally) (Colloredo-Mansfeld, 1998); and 5) *ventriloquism* (Guerrero, 1997).[7] We provide extended descriptions for each manifestation in Appendix B. Since we are not interested in independently coding these different manifestations but in identifying covert racist language more broadly, we are not concerned with slight overlaps in the definitions.[8] In Appendix B, we also include examples of covert and overt racist language from our corpus.

[7] On page 7, we identify five generalizable characteristics of covert discourse. Some of the manifestations of covert racism in the Ecuadorian case fit within this description. For example, no-difference racism or the negation of the indígena identity minimizes the experiences of member of the the indígena community, going even a step further by recriminating member of the community for "making it all about race" when they are all "are Ecuadorians". Or the understanding of "hygienic racism" is contingent on understanding social codes about the racist history describing the indígena population. There are other manifestations that have less generalizable characteristics. For example, in addition to Ecuador, Murrey and Jackson (2020) find patterns of ventriloquism when describing oil development and *localwashing* in Central Africa (i.e., oil companies speaking on behalf of communities and supporting extractive practices).

[8] For our example, overlaps avoid gaps in the general identification of covert and overt racism. However, researchers interested in coding each manifestation separately should make the definitions of the different concepts distinct and exclusive.

2.3 Data

To build our training set, we use Twitter data covering indígena-related discourse in Ecuador. The main corpus of tweets was collected during the indigenous protests in Ecuador between October 1st and October 30th, 2019.[9] The corpus includes 2,020,487 posts (168,933 unique posts) by 91,458 unique Twitter users. A second corpus of tweets includes posts mentioning the indígena community in Ecuador between 2018 and 2021.[10]. The corpus includes 1,497,369 posts (154,630 unique posts) by 66,574 unique Twitter users. This second corpus allows us to test generalizability and control for inter-temporal losses in accuracy.

We used 3,724 tweets to create the training set. We sampled the corpus using three different approaches to have a more balanced training set. The first third of the training set was randomly produced from the entirety of the corpus. Another third of the training set was randomly produced from tweets that included the word 'indígena' or 'indio.'[11] The final third of the training set was randomly produced from tweets that contained linguistic markers (i.e., words or phrases) that are commonly associated with the indígena community, including those that directly pertain to the indigenous identity (see Table B.1). In Table 1, we provide examples from the data for each form of racist discourse previously described.

2.4 Codebook, Labeling, and Inter-Coder Reliability

The complete codebook from our example used to train coders is available in Appendix C. The codebook structure took the following form: 1) a general definition of covert and overt racism, 2) rules, explanation of rules, and examples of overt racism, 3) rules, explanation of rules, and examples of covert racism, 4) advice for coders to handle unclear/ambiguous text. To generalize this codebook, we encourage

[9] We collected the data by connecting *rtweet* to Twitter's application programming interface (API). We used the following terms in the search: *paro* and *ecuador*.

[10] We used the following term in the search: *conaie*, *indio ecua*, *protesta ecua*, *indígena ecua*, *mestizo ecua*. 'CONAIE' is the Confederation of Indigenous Nationalities of Ecuador or *Confederación de Nacionalidades indígenas del Ecuador*, the largest indigenous organization in Ecuador; 'protesta' is strike in Spanish; and 'mestizo' (translated to 'mixed person') is the most common ethno-racial identity in Ecuador

[11] Another term used to refer to indigenous peoples is "indios," which carries derogatory connotations when used by the blanco-mestizo population. Despite efforts by the indigenous community to reclaim the term, it often serves as a racial slur. However, it's important to note that the use of "indio" does not always indicate racism.

Covert Racism	
Type	**Tweet (example)**
No-difference racism	“This MESTIZO [sic] just like all ecuadorians is called CARLOS PEREZ, who disguises as indígena y makes people call him Yaku, jah!!”
Attacks on the capabilities	“The [CONAIE] is a threat to the progress of the indigena population. Instead of destroying, focus on building a better country. Focus on educating and getting them out of their ignorance.”
Infantilization	“It is enough for Rafael Correa @MashiRafael to tweet his messages. It says a lot about you and the movement your lead @jaimevargasnae. They use the indigenas only to benefit the leaders, that is to say you!”
Hygienic racism	“we will play carnaval* with the indígenas, they fear water.”
Ventriloquism	“At this point this is not a strike for the economic policy. This is to destabilize, they are affecting the infrastructure of the State. The indigenas are stooges, they will not negotiate, they want to overthrow @Lenin.”

Overt Racism	
Type	**Tweet (example)**
Ethnic slurs	“¡¡¡Damned *Longo*!!! ¡you will not come back!”
Attacks explicitly mentioning the ethnic identity	“This is one is stupid even when they fix their stupidity. Who told him it is a country inside of another country? ¡Stupid indio!”

Table 1: Examples of racism in Ecuadorian Twitter. All tweets are translated from Spanish. See Table B.2 for the original text. Note: ‘CONAIE’ is the Confederation of Indigenous Nationalities of Ecuador or *Confederación de Nacionalidades indígenas del Ecuador*, the largest indigenous organization in Ecuador, and ‘mestizo’ (translated to ‘mixed person’) is the most common ethno-racial identity in Ecuador.

researchers and practitioners to maintain a similar structure: a general definition of the different forms of racism to be coded, followed by the various manifestations of those forms that coders might encounter in the training set, and advice on how to handle unclear or ambiguous text. The codebook structure allows researchers to label different forms of racist discourse while providing important details of each category that help coders have a more homogeneous understanding of each element. The final form of our codebook resulted from numerous training and discussion sessions with coders. One of the greatest difficulties we found in this process was labeling discourse that is, by nature, covert (e.g., covert racism, symbolic racism, laissez-faire racism, etc.). Thus, in Appendix C, we add details on training coders, lessons learned, and suggestions on tackling drawbacks found in the process.

For this particular exercise, we followed Schreier (2012) in the coding process. We trained two hired

Table 2: Training Data by Type of Tweet

Type	N	%
Non-racist	2,187	58.7
Covert	1,035	27.8
Overt	501	13.5
Total	3,723	100

coders across three review rounds to ensure consistency and performance. The initial review round (500 tweets) allowed us to explain discrepancies and identify cases we did not initially consider. After a second review round (500 tweets), coders independently coded 2750 tweets (Cohen score = 0.9 or strong agreement). In Appendix C, we provide details on revisions to the codebook to avoid ambiguities, common questions from the coders, and lessons learned throughout the process that can be generalized to other cases and can be helpful to researchers.

To harmonize our training set when encountering coding discrepancies, we coded tweets as covertly racist or overtly racist if both coders agree. In Table 2, we present the distribution of categories in our training data. Notice that the data is skewed in favor of non-racist tweets, which is expected. Racism, while not uncommon, is not the main lexical form found in social media, as it is, we argue, a costly social behavior that can have real-life consequences. It is also monitored, flagged, and eliminated from most social media platforms.[12]

3 An NLP Approach to Detecting Overt and Covert Racism

Once we have a theoretically-driven training set, we turn to our NLP approach to classifying overt and covert racist discourse and apply it on Twitter corpus.

There is an expansive literature advancing the detection of explicit hate speech and racism in language (see Schmidt and Wiegand, 2017; Gambäck and Sikdar, 2017; Ousidhoum et al., 2019; Yoder et al., 2022). Extant work on NLP approaches for detecting implicit hate speech in large corpora have highlighted the challenged of the task, pointing to the linguistic nuances and diversity of the implicit hate class

[12] However, Twitter's automatic detection of hate speech in Spanish is rather lax.

(ElSherief et al., 2021), their failure to generalize to other corpora (Kim, Park and Han, 2022; Kim et al., 2023), as well as the lack of clear definitions for concepts like 'implicit and explicit hate speech' (Ocampo et al., 2023).

Some of the computational work to detect covert racist language in large corpora focuses primarily on improving the performance of models by modifying the models (e.g., different fine-tuning strategies, pre-training approaches) without engaging with the source data (e.g, training set) or how the model can adapt to characteristics of the concepts analyzed. For example, Gao, Kuppersmith and Huang (2017) use an explicit slur-term learner and a neural net classifier (i.e., LSTM) to capture explicit and implicit hate speech. Kim et al. (2023) use machine-generated data and contrastive learning to improve the performance of a pre-trained model to detect implicit hate speech, an approach that is complemented by Ahn et al. (2024) who leverage the shared semantics among the data.

The approach that more closely resembling ours is proposed by ElSherief et al. (2021). They develop a taxonomy of implicit hate speech that include elements such as incitement to violence, irony, and inferiority language.[13] They then test their data set to train several variations of SVM and BERT models (e.g., SVM + TF-IDF, BERT + ConceptNet). In addition to providing a complete workflow (i.e., from the theoretical foundations to create a training set to the model choice), we explicitly link the characteristics of racist discourse with the mechanics of machine-learning models, mainly Transformers-based models, to obtain improved performance.

In the following section, we describe Transformers, a deep learning architecture used for text classification developed by Google in 2017 (Vaswani et al., 2017). We also explain how to exploit the full functionality of XLM-RoBERTa, the Transformers-based model we train to detect racism in text.[14] We briefly introduce the most common NLP approaches in the social sciences, which we use as baselines to compare to our main model. We then describe our process for further pre-training and fine-tuning the XLM-R model to detect overt and covert racism specifically in the Ecuadorian context. Throughout this

[13] It is unclear how ElSherief et al. (2021) decide on the elements that make up the taxonomy, or the code book used by annotators.

[14] We use the cross-lingual version of RoBERTA because our text is in a non-English language.

section, we use the Twitter data corpus from Ecuador described in the previous section.

3.1 Pre-trained Contextual and Non-Contextual Embeddings

Previous work in the social sciences has shown the benefits of word embeddings in identifying the meaning of words in context. Word embeddings represented a major leap forward in NLP and have advanced many substantive debates through novel and sophisticated analyses of text. For example, Rheault and Cochrane (2020) combine word embeddings with political metadata to train large-scale text from global Parliaments and produce scaling models for ideological placement. Likewise, Rodriguez and Spirling (2022) propose a method, based on Khodak et al. (2018), that utilizes a small sample of tokens of interest to estimate new embeddings that are able to capture changes over time, like partisan identity, or some other document-level covariates. We extend the intuition behind both approaches in our application of Transformer-based models. Using the theoretical insights from Steps 1 - 3, we create custom embeddings for highly informative words that are not included in the pre-trained model (e.g., slang) and further train our model using data from the target context, before fine-tuning the model for our specific task. In other words, we take a pre-trained model that in its base form outperforms traditional word-embedding models, and adapt it to our specific context to maximize performance.

Our model of choice is XLM-RoBERTa, or XLM-R, the cross-lingual version of RoBERTa (Liu et al., 2019; Conneau et al., 2019).[15]. The model is based on the Transformers architecture, as are other well-known models such as BERT and DeBERTa.[16] Transformers are large neural networks that produce representations (i.e., embeddings) of text input through the self-attention mechanism (Vaswani et al., 2017). Self-attention relates each word to all other words in the sentence, which makes BERT and similar models *bidirectional* in nature. Unlike other sequential and unidirectional methods such as Word2Vec, Transformers-based models can process tokens in a sentence *all at once*, which makes the em-

[15] We use a version of XLM-R that is further pretrained in Spanish named entity recognition. The model's name is 'xlm-roberta-large-finetuned-conll02-spanish' and can be found at `huggingface.co/FacebookAI/xlm-roberta-large-finetuned-conll02-spanish`

[16] For more detail on Transformers models and the performance differences across Transformers models, see Timoneda and Vallejo Vera (2025*a*).

beddings inherently dynamic across contexts. This adaptability gives Transformers models unparalleled performance in supervised classification with highly nuanced and complex text.

XLM-R is trained on large amounts of pre-existing text. The creators of XLM-R used 2.5 terabytes of filtered data from Common Crawl and it contains 250,002 unique vocabulary elements (Tunstall, von Werra and Wolf, 2022).[17] For words that do not directly match a token, XLM-R adds different word chunks (tokens) together and extracts a combined representation. For instance, the word 'training' would be tokenized as "train, ##ing", with the double hashtag indicating that 'ing' is a subword token that follows the token 'train'. Subword tokenization strategies, or out-of-vocabulary strategies, have proven quite accurate at handling unknown words. However, if a word is important enough to a researcher's specific application, XLM-R can be further trained to understand this word in context. In this article, we detail how this process to further train a Transformers model works and leverage the flexibility of XLM-RoBERTa to improve its recognition of racist text in Spanish (Timoneda and Vallejo Vera, 2025*a*).

4 Training an XLM-R Model to Detect Overt and Covert Racism

In all supervised machine learning models, there are three important parts to consider: the input text or training data, the model to train or fine-tune, and the testing and validation of the results. We place emphasis on all three steps, showing the importance of (1) building a theoretically-grounded and rigorous training set, (2) pre-training and fine-tuning an XLM-R model built specifically to detect racism in Ecuador, and (3) applying 10-fold cross-validation and reporting accurate out-of-sample performance averages for model testing. In the previous sections, we describe the first step. Below we go over steps 2 and 3 in detail and present the main results of the article.

[17] Common Crawl is a nonprofit organization that scrapes websites and stores all their text and shareable information.

4.1 Pre-training the XLM-R for the Ecuadorian Context

Timoneda and Vallejo Vera (2025*a*) show that further training a Transformers model such as XLM-R can yield significant increases in performance. There are four steps to further train an XLM-R model. First, we add new tokens to the original XLM-R tokenizer that reflect different expressions of racism in Ecuador. Second, we assign the mean representation of similar words to the newly added tokens. Third, we feed a new large text corpus containing the new tokens and train it again to improve the representations for those tokens.[18] Fourth, we save the new model and apply it to our classification task through fine-tuning in the same way we would apply the original.

We followed these four steps to build our own XLM-R classifier, 'XLM-R-Racismo'. We first add 20 tokens to the XLM-R tokenizer, increasing the number of tokens in the vocabulary (tokenizer) to 250,022. We produced the list of 20 tokens based on our knowledge of the Ecuadorian context and what we learned while labeling our training data. Specifically, we found a series of terms that strongly signaled overt and covert forms of racism and shorthand expressions used, in part, to avoid being flagged as inappropriate content by Twitter (e.g., instead of 'hijo de p*ta', users would write 'hdp'). These words were either not in the pre-trained vocabulary or appeared in an unrelated context (e.g., "longo" is a derogatory term in Ecuador, yet it only appears in the pre-trained data as the Portuguese word for "long"). The subword tokenization technique that XLM-R uses by default would also not help understand the meaning of 'hdp' in context. Thus, we added tokens such as *hdp*, *longo*, *guangudo*, *poncho dorado*, and *cholo*, among others, to the tokenizer.[19]

Tokens added to the vocabulary are given the mean embedding for similar preexisting derogatory terms.[20] Doing this imbues these tokens with initial substantive meaning, which helps the model recognize their contextual relationship to other tokens during training.

[18] Note that training for this third step is unsupervised.

[19] See Appendix E for a complete list.

[20] Similar terms are synonyms or words used for a similar purpose. For example, when adding insults to the vocabulary, we give them the mean embedding of other insults. Thus, we provide to the added token 'mamaverga' the average embeddings of 'f*cker', 'idiot', 'd*ck', and 'stupid', all insults used in similar contexts. For a complete list of added tokens and their initial embeddings, see Appendix E.

Afterwards, we use a new unlabeled set of over 1.7 million tweets to pretrain the model on new text.[21] The new corpus contains the added tokens and allows the model to see how and where these words are used in context. The unlabeled Twitter corpus also provides additional information on the unique linguistic construction of tweets. During this process of further training, the embeddings for the new tokens will change, making them a more accurate representation of their actual meaning. Furthermore, the embeddings of tokens that rarely appear in the original data but that are used often in the new unlabeled corpus will also change and improve. Therefore, both adding new tokens and further training the model on new unsupervised text combine to maximize model performance for highly specific tasks. Through further training, we embed the newly added tokens with unique meaning based on the contextual characteristics of the relevant corpus.

Once further training is complete, the resulting new model is saved and applied to text in the same way as the original XLM-R model. Importantly, this process can be extrapolated to other research questions, especially those in a very specialized fields for which the original XLM-R model does not have a strong set of pretraining text. XLM-R-Racismo, whose full technical model name is 'xlm-r-racismo-es-v2', is available for public use at `huggingface.co`. This is the model we apply below and the one to which we compare the performance of all models, including the original XLM-R.

4.2 Fine-tuning our Models

We fine-tune two variants of our main XLM-R model: (1) the original XLM-R model[22] and (2) our own XLM-R model pretrained specifically to detect racist text with Twitter data from Ecuador.[23] To

[21] For the unlabeled set we downloaded tweets that included the following terms: *indio*, *indígena*, *longo*, *guangudo*, and *yunda*. The set covers tweets produced between January 2018 and December 2020.

[22] The version we use has been further pretrained for named entity recognition in Spanish.

[23] Before we fine-tune our models, we need to specify values for the hyperparameters. Following the recommendations by the authors of BERT and XLM-R and our own cross-validation tests, we use a batch size of 32, a learning rate of 2e-5, a maximum sequence length of 85, and 4 epochs. Finally, we use the weighted Adam optimizer (Loshchilov and Hutter, 2017), which performs well for NLP data and for BERT and RoBERTa models –and their cross-lingual variants– in particular. Note that Batch size refers to the amount of tokens that the algorithm will analyze simultaneously. For BERT, 16 or 32 are recommended for highest accuracy (Liu et al., 2019). Learning rate is the size of the steps used by the algorithm in each iteration towards a minimum of a loss function. The recommended learning rate for a BERT or RoBERTa model is $3e^{-5}$, however we modified the learning rate slightly to adapt it to our training set. The number of epochs is the number of times the model runs through the entirety of the data. For BERT and RoBERTa, the recommended number of epochs

evaluate model performance, we use 10-times repeated cross-validation and report the usual metrics for NLP tasks, including accuracy, recall, precision, and F1-scores for each model. We are particularly interested in understanding how our model is misidentifying covertly and overtly racist language.[24]

We also train six separate models to serve as baselines, all of which are widely used in the social sciences and have been considered state-of-the-art at some point or another in the past decade (Grimmer, Roberts and Stewart, 2022). First are non-contextual word-embedding models such as GloVe and Word2Vec, which are usually applied via Convolutional or Recurrent Neural Network architectures (CNN or RNN). For this exercise, we use a CNN with Word2Vec and a Long-Short Term Memory RNN with GloVe embeddings. Bidirectional Long-Short Term Memory (Bi-LSTM) networks are a type of RNN that have proven particularly effective with NLP tasks due to their capacity to retain some contextual information from previous word sequences in the text. We are thus especially interested in comparing XLM-R's performance with that of a Bi-LSTM neural network using GloVe embeddings in the context of covert racism in Ecuador.

Additionally, we use two generative models to provide a performance comparison between them and our main RoBERTa models. Recent research has explored the use of generative models to detect hate-speech, focusing on the potential benefits as an annotator, rather than a trained model to classify text at scale (Huang, Kwak and An, 2023; Kruk et al., 2024*b*). For this test, we use both OpenAI's GPT-4o and Meta's Llama 3.3-70B as annotators, following Timoneda and Vallejo Vera (2025*b*).[25]. For both models we follow a few-shot approach with chain-of-thought (CoT) reasoning to maximize performance (see Gilardi, Alizadeh and Kubli, 2023).[26] Lastly, we also train two traditional machine-learning models

is between 2 and 4. The maximum sequence length is the number of token at which we truncate each observation (i.e., tweet). We do this to manage computer memory. We use 10-fold cross-validation for all models and report the averages.

[24] Depending on the application of this method, researchers might be more interested in the proportion of false negatives or false positives. This will depend on the ultimate goal of the project. We encourage researchers to pay particular attention to the possible effects of having more (or fewer) false positives or negatives on the conclusions obtained from estimations using these predictions.

[25] Meta's Llama family is an open-source and free suite of LLMs, an important element when considering our models of choice (Palmer, Smith and Spirling, 2024). Models from the Llama family are fine-tuned and then improved by combining techniques that use preference ranking to signal to the model the best response from several LLM responses (Wu et al., 2024)

[26] CoT reasoning combines both examples and their true labels with a reasoning paragraph detailing why the sentence was classified the way it was. It has been shown to produce best performance with few-shot learning (see Gilardi, Alizadeh

to provide two non-neural network baselines: Support Vector Machine (SVM) and Logistic Regression (LR). For these two models, we tokenize the words using NLTK's Spanish word tokenizer and use a TF-IDF vectorizer. We then train and test the models, and use 10-fold cross-validation to evaluate model performance and results (Loper and Bird, 2002).

4.3 Results and Validation

The performance gains from the XLM-R models are considerable compared to the traditional machine learning models and CNN and Bi-LSTM using GloVe and Word2Vec embeddings. These results are in Table 3. This is especially true with covert racism, which is much more difficult to detect without deep contextual understanding of text. Both the logistic regression and SVM models score below 50%. The CNN and Bi-LSTM models improve the classification for covert racism, as expected, given the more advanced word embeddings used and their greater capacity to understand context. The Bi-LSTM model correctly detects covert racism more than 50% of the time. However, the two XLM-R models in Table 3 significantly outperform the rest of the models. The original XLM-R model identifies covert racism correctly 73.3% (0.733) of the time in unseen data. Even more striking are the results of the XLM-R-Racismo model further pretrained on 1.7 million racist tweets from Ecuador (last row in Table 3). The model shows strong improvement in covert racism with an F1 score of 0.805. This represents a 9.56% improvement in performance when compared to the original XLM-R model and 41.5% when compared to the Bi-LSTM model. As for the generative models, their performance is significantly below XLM-R, especially in terms of covert and overt racism. GPT-4o with few-shot CoT produces F1 scores of 0.624 for covert and 0.483 for overt racism, while Llama's scores are lower at 0.445 and 0.416, respectively. This suggests that while these models are trained on even greater amounts of data than XLM-R, they do not perform as well in highly specific cross-lingual tasks.

The results are less pronounced, as expected, with overt and non-racist text. SVM and Logistic

and Kubli, 2023).

Table 3: Performance statistics for racist discourse classifiers (10-fold CV averages)

Model	**Non-Racist**				**Covert Racism**				**Overt Racism**				**Macro-Average**			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
ML - Logistic Reg.	**0.963**	0.744	**0.963**	0.839	0.300	0.651	0.300	0.407	0.649	**0.892**	0.649	0.748	0.756	0.637	0.762	0.665
ML - SVM	0.925	0.808	0.925	0.862	0.451	0.611	0.451	0.514	0.746	0.849	0.746	0.789	0.783	0.708	0.756	0.722
CNN - Word2Vec	0.905	0.791	0.905	0.843	0.484	0.625	0.484	0.528	0.615	0.820	0.615	0.697	0.756	0.668	0.745	0.689
Bi-LSTM - GloVe	0.877	0.841	0.877	0.857	0.568	0.590	0.568	0.569	0.745	0.812	0.745	0.775	0.780	0.730	0.747	0.733
GPT-4o F-S CoT	0.755	0.908	0.755	0.825	0.624	0.562	0.624	0.592	0.605	0.402	0.605	0.483	0.699	0.744	0.699	0.714
Llama 3.3 F-S CoT	0.578	**0.934**	0.578	0.714	0.445	0.390	0.445	0.416	0.701	0.296	0.701	0.416	0.558	0.696	0.558	0.591
XML-R	0.865	0.890	0.865	0.876	0.759	0.712	0.759	0.733	0.761	0.788	0.761	0.771	0.818	0.795	0.797	0.794
XML-R-Racismo	0.904	0.927	0.904	**0.915**	**0.817**	**0.797**	**0.817**	**0.805**	**0.834**	0.808	**0.834**	**0.818**	**0.851**	**0.803**	**0.844**	**0.846**

Regression models perform fairly well in identifying overtly racist and non-racist discourse, with accuracy scores over 90% for the latter and over 60 and 70%, respectively, for the former. Bi-LSTM performs well with overt racism and non-racist discourse, with an accuracy of 77.5% and 85.7% respectively. Original XLM-R, for its part, identifies overt racism correctly 77.1% (F1 score) of the time and non-racist discourse in 87.6% of instances. XLM-R-Racismo still improves on these figures, with an accuracy of 91.5% for non-racist discourse and 81.8% for overtly racist discourse. These numbers represent a 5% average gain over XLM-R and 6% over Bi-LSTM in these two categories. Taken together, this is a substantial improvement in detecting covert and overt racism (especially covert), considering we only used a relatively small corpus of 1.7 million tweets for pretraining, which pales compared to the 2.5TB of data used to pretrain the original XLM-R model. It is reasonable to think that larger amounts of task-specific pretraining data would further improve model performance during fine-tuning.

While it is expected for Transformer-based model to outperform other machine-learning approaches, the results are particularly noteworthy given that similar work on detecting hate speech in Spanish text has achieved F1 scores for the hate-speech category of 72.7% (Gertner, 2019) and 75.5% (Plaza-del Arco et al., 2021), compared to our 81.8% F1 score for the overt racism category. Both of these works used similar Transformers-based infrastructures based on Google AI's original models –specifically, they used mBERT (multilingual BERT) and BETO, a refined BERT model pretrained specifically with large amounts of text in Spanish. The performance gains of the XLM-R-Racismo are substantial compared to these two works (12.52% and 8.34% respectively). There is no comparable extant study of *covert* racism, so our figures showing F1 scores over 0.8 in this category show the true potential of our approach to identify highly nuanced categories in text.

Beyond the performance of our models, we are interested in conducting error analysis to identify the weaknesses of our best-performing set-ups. For the XLM-R-Racismo model, we check each mislabeled instance and compare instances that were commonly misclassified. Since we have a multi-class classification model, it is not sufficient to analyze the false positive (FP) or false negative (FN) rate, as it is not the same to mistakenly label a non-racist text as covertly racist and to label an overtly racist text as covertly racist. Thus, not only are we interested in identifying the FP and FN rate for each category, but more importantly, we are interested in identifying towards which category is the error skewing.

		True Category		
		Non-racist	Covert racism	Overt racism
Predicted Category	Non-racist	**188**	26	8
	Covert	22	**53**	9
	Overt	3	7	**70**

Table 4: Confusion matrix for the pretrained XLM-R model

Table 4 shows the confusion matrix for the XLM-R-Racismo model. Our model is accurate at predicting overtly racist and non-racist text. In the model, text that is misidentified as overtly racist is more likely to be covertly racist (8.8%) rather than non-racist (3.8%). This is a positive result, as covertly and overtly racist text is, above all, racist text (just different manifestations of the phenomenon). As is evident from all models, covert racism is more difficult to identify. XLM-R-Racismo is more likely to misidentify covertly racist text as non-racist (26.2%) than overtly racist (10.7%), which is consistent with our expectations. More important, however, is the relatively low FP rate for non-racist text. The XLM-R-Racismo model misidentifies (15.3%) of non-racist text was classified as either covertly or overtly racist.

Having shown the advantages and limitations of our approach, we use our trained model to predict racist discourse in a corpus of 168,933 unique tweets. Following the same rules used to label our training data set, we manually code a sample from the predicted data. We are interested in the extent to which our model is able to scale up to new text. We present the performance statistics in Table 5. Compared to the results from Table 3, this out-of-sample performance is similar to the ones obtained with the training data.

		True Category		
		Non-racist	Covert racism	Overt racism
Predicted Category	Non-racist	**214**	1	0
	Covert	40	**92**	1
	Overt	21	10	**96**

Table 5: Confusion matrix for XLM-R-Racismo predictions and hand-coded out-of-sample data

5 Covert and Overt Racism in Ecuadorian Twitter

To showcase the usefulness of our method, we apply our trained model to predict racist discourse on the full Twitter dataset. We explore the prevalence of covert and overt racism in the time covered by the data, and the characteristics of the users that are more likely to engage with the different forms of racist discourse. We analyze the relationship between the use of racists language and the importance of the user within the network, as well as their likely position towards the indígena protest. In addition to the abundance of text, this corpus includes information (i.e., tweets) from public figures and less-prominent users during the political turmoil that highlighted the historical confrontation between the indígena and the blanco-mestizo community (Vallejo Vera, 2023). As previously mentioned, we use our trained model to predict racist discourse on the two corpora described in the Data section: Twitter data during the 2019 indígena Protest in Ecuador (Protest Data), and Twitter data mentioning the indígena community between 2018 and 2022 (indígena Data). In Table 6, we show the distribution of non-racist, covertly racist, and overtly racist tweets in both corpora.

Table 6: Prevalence of Different Forms of Racist Tweets

	No Racism	**Covert Racism**	**Overt Racism**	**Total**
Corpus Protest Data	159,314	7,312	1,436	168,062
Corpus Indigena Data	144,520	14,350	3,315	153,769

Social media users tend to cluster around like-minded peers, leading to homogeneous communities–clusters of nodes where the same information (tweets) is shared–that are consistent across time (Calvo and Aruguete, 2 On Twitter, communities formed around political events and cleavages often have at their center political leaders or users strongly aligned with the leadership. In Ecuador, this roughly translates into pro-government users mostly interacting (retweeting) with other pro-government users or pro-indígena users interacting primarily with other pro-indígena users (Vallejo Vera, 2023). We identify communities in our Twitter data via a random-walk community detection algorithm (Pons and Latapy, 2005),where each user is a node that creates a connection (edge) when they retweet another user (node).[27] The random-walk community detection algorithm identified the same primary communities in both datasets: a pro-government network, which includes 41,493 nodes in the Protest Data and 26,036 nodes in the Indígena Data; and an indígena community network of 30,244 nodes in the Protest Data and 26,110 in the Indígena Data. The community detection is, in essence, an unsupervised learning problem based on the characteristics of a network. To determine which type of users are part of each community, we look at highly influential users (i.e., users with high in-degree) in each cluster: the pro-government community had at its center then-President @lenin, then-vice-President @ottosonnenh, and then-interior minister @mariapaularomo; in the center of the pro-indígena community was the institutional account of the @CONAIE_Ecuador, and its then-president, @jaimevargasnae.[28]

A correlational analysis of our data aligns with the theoretical expectations from our review of covert and overt racism. For example, Bonilla-Silva (2015) suggests that covert racism has replaced the overt manifestations of the racist ideology, partly as an answer to demands to dismantle the racial structure without actually changing it. Furthermore, unlike covert racism, overt racism is socially punished (i.e.,

[27] We provide more details on how we create the networks and the choice of our community detection algorithm in Appendix F.

[28] While community-detection algorithms have been validated in network analysis (for Twitter, see Aruguete and Calvo, 2018), any community-detection algorithm can cluster within a community users who might not align (ideologically, culturally, or socially) with most users in that community. Given the size of the data, it is difficult to manually verify every individual node. The results from our analysis align with the expected behavior (e.g., users in the pro-government community were more likely to engage with racist discourse than user in the pro-indígena community). We also show that some of the unexpected behavior (e.g., users in the pro-indígena community engaging in overtly racist behavior) was partly due to misclassification of the algorithm.

a socially costly behavior). We find, for example, that covert racism is more prevalent than overt racism in both corpora. Given the socially costly nature of overt racism, we also find that verified accounts–political figures and media accounts–do not produce or reproduce overtly racist content. The stakes for public figures in engaging with racist content, in terms of exposure and repercussions, far outweigh those for less public users. Verified users predominantly include news agencies and government officials, who face greater risks in publicly endorsing overt racist discourse.

To more formally show this relationship, we regress a multinomial logistical model with a categorical dependent variable for whether the tweet is non-racist, covertly racist, or overtly racist, against the (log) in-degree of the user tweeting, as well as the community to where they belong (i.e., pro-indígena community or pro-government community). While users at the center of the network topography are less likely than less prominent users to tweet covertly racist content, they do so at a higher rate than overtly racist content. The lower cost of these discursive forms allows more prominent and more public users –users whose reputation would suffer from engaging with overtly racist content– to produce and reproduce it. This includes members of the media (e.g., online political magazine @4pelagatos), politicians (e.g., former-president @mashirafael and then-president @lenin), and political commentators (e.g., pro-government journalist @CarlosVeraReal). For example, a common characterization of the indígena community during the protests was as "terrorists" or "criminals" (one of the characteristics of "hygienic racism"). Then-President Lenin Moreno tweeted: "The indígena leadership, supportive of Democracy and the Rule of Law for all Ecuadorians, must stir away from these false leaders, akin to indigenous terrorism and paramilitarism."

The opposite is also true: less influential users (i.e., users with low in-degree) and users in the periphery of the network topography produce most of the overtly racist content. We plot the predicted probabilities in Figure 1 for ease of interpretation. The results suggest that, as in-degree increases, the probability of tweeting racist content decreases. For example, going from the first quartile of the in-degree distribution to the third quartile reduced the probability of producing racist content by half.

Furthermore, research has shown how bots–social media accounts controlled by software–amplify hate speech to sow discord and perpetuate harm (see, for exmaple Uyheng, Bellutta and Carley, 2022). To explore how bots engage with racist content, we follow Yang et al. (2020) and estimate the likelihood a user is a bot. The method proposed by Yang et al. (2020) yields a score between 0 and 1, where 0 is behavior akin to human users and 1 is more bot-like behavior.[29] We find that, as suggested, a higher bot-like behavior is positively associated with the likelihood of reproducing covert and overt racism. Going from no bot-like behavior to a a high bot-like behavior ($\geq .5$) increase the likelihood of producing covert racism from 4% to 8%; for overt racism it increases from 1% to 2%.

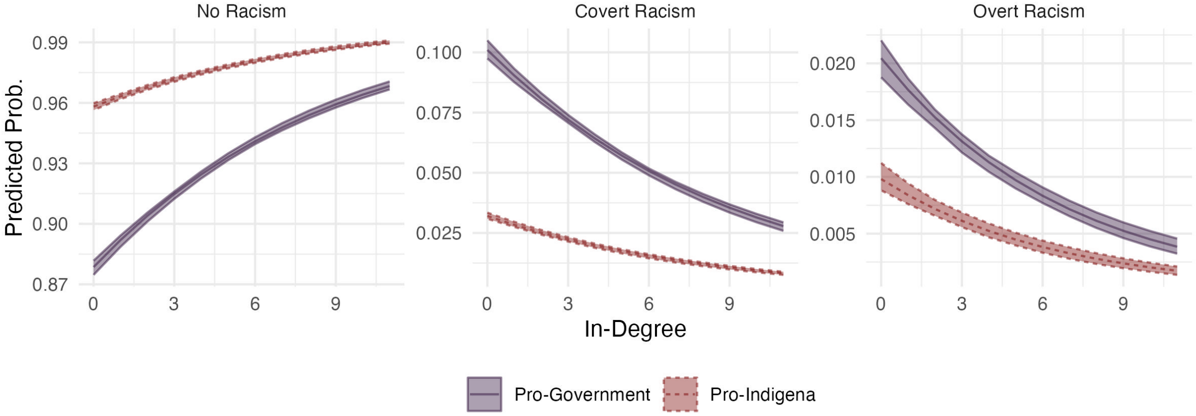

Figure 1: Predicted probabilities of tweeting covertly, overtly, or none racist content during the Ecuadorian protests (2019). We find a similar result using the data from the network mentioning the indígena community (2018-2022).

The difference in the prevalence of covertly and overtly racist content between pro-government and pro-indígena communities speak to the literature on racism and on social media. For the former, the presence of covertly racist content across both communities points to the normalized nature of covertly racist language in Ecuador. Users from the pro-government community retweeted 77.2% of all covertly racist content (37,025 retweets), while users from the indígena community retweeted 18.8% (9,076 retweets).

[29] Using the trained model proposed by Yang et al. (2020) requires a connection to their API and access is limited by the number of requests users are able to make. Given this, we sample 20,000 observations (tweets) from our data and use the API to predict the likelihood a user has bot-like behavior. We then estimate a similar multinomial logistical model than the original, this time for the sample, with a categorical dependent variable for whether the tweet is non-racist, covertly racist, or overtly racist, against the (log) in-degree of the user tweeting, the community to where they belong (i.e., pro-indígena community or pro-government community), and the likelihood a user has bot-like behavior. We show the complete model in Appendix G.

Even though it is impossible to individually know the racial identity of each user, the participation of mestizos in social media platforms corresponds to their distribution in the population (LAPOP Lab 2019). Research has shown that for mestizos in Ecuador (Beck, Mijeski and Stark, 2011; Traverso-Yépez, 2005), as for whites in the U.S. (Bonilla-Silva, 2013) and Europe (Coates, 2008), socializing in and benefiting from the racist social order can lead them to reproduce covertly racist language to rationalize racial inequalities, hierarchies, and behaviors. Furthermore, some covert forms of racism are embedded in apparently benign intentions (see Section 4). Covert forms of racism from supporters of the indígena community often perpetuate the idea of indígenas as a cursed race, a people in need of protection (i.e., *infantilization*), accused indígena leaders that did not support the strike of not being "real" indígenas (i.e., *denying their identity*), or questioned whether those leaders involved in the strike were 'real' indígenas. For example, a user supporting the strike tweets: "Yaku Pérez [an indígena leader] is a fake indígena, trying to [take advantage of the situation]. CONAIE cannot allow for this treason."

The distribution of overtly racist content across both communities, on the other hand, speaks to behavior of users in social media. Of all overtly racist content in the Protest Data (6,446 retweets), users from the pro-government retweeted 75.4% (n=5,065), while users from the pro-indígena community retweeted 20.6% (n=1,128). Arguably, we should find no overtly racist behavior the pro-indígena community. We find that some of the overtly racist content from the indígena community were wrongly predicted tweets (35.5%).[30] The rest were tweets produced by a limited number of users (n=233) that were, arguably, wrongly categorized nodes, or nodes whose purpose is to disrupt online communities (e.g., bots or trolls). The average bot-score for user in the pro-indígena community is of 0.18, compared to 0.14 in the the pro-government community; 5% of users in the pro-indígena community had a high likelihood (bot-score ≥ 0.5), compared to 3% of user in the pro-government community. Research has found a positive correlation between the presence of bots and of trolls (i.e., human-controlled accounts that spread,

[30] Most of the wrongly predicted tweets in this corpus were texts where it was not clear to whom a particular insult was directed. While insulting somebody is not a racist act, doing so by priming their identity is. For example, in the tweet: 'CONAIE [sic] attacking the wrong person. There are incompetent police officers, just as idiotic as the thugs.' it is not clear if the user is calling 'idiotic' and 'thug' the CONAIE, which is a reference to the indígena identity. Our model in these cases assumes that the user is, in fact, referring to the indígena identity.

among other things, hateful messages), suggesting that some of the users are also more likely to be trolls. Despite not being users supporting the indígena community, that they were included in their online community suggest that (some) pro-indígena users were likely exposed to these messages, and their pernicious effects, including consequences to mental health (Saha, Chandrasekharan and De Choudhury, 2019) and engagement (Gagliardone et al., 2016).

6 Limitations

While our proposed Transformers-based approach to classifying racist text improves upon the performance of other supervised machine-learning models, it is not without its limitations. First, while not uncommon, overt and covert racist discourse is not the main lexical form found in "naturally-occurring" corpora, as producing racist text is socially costly. This can lead to highly unbalanced samples (or require large samples for labeling). Unbalanced training data sets produce less accurate results than balanced data sets, particularly when using small samples to train models. Research has shown that the transformers architecture used in this paper outperforms other supervised-learning approaches (e.g., CNNs and RNNs) with small samples. Yet, it does not guarantee a baseline level of accuracy that researchers might be comfortable with. Researchers might use dictionary-based techniques to select the training set to address this limitation. Certain words might signal a higher likelihood of observing the phenomenon of interest –in our case, racist discourse.

Second, as social scientists, we are usually comfortable with error, as long as it does not overestimate our predictions (type-I error). When examining the performance of our model classifying overt and covert racism, it more often mislabeled racist discourse as non-racist than non-racist discourse as racist. That is, the model produced more false negatives than false positives. This is generally preferable over type-I error, and we can argue in our analyses that we are underestimating the effect size rather than inducing positive bias. However, this might not be the case for practitioners in other fields. For example, when applying our method to flagging potentially harmful content in social media comment sections, researchers might want to favor an over-cautious algorithm that is more likely to flag non-racist content as racist to have a human later decide on the accuracy of the prediction.

Third, to identify certain instances of racist discourse, specific *knowledge* of the case is required. For example, following our codebook, this tweet should be classified as covertly racist: “I will stop being polite to this deceitful mestizo, Carlos Perez you are an accomplice to the damage done to my Ecuador.”[31] However, to recognize the racist nature of the tweet (i.e., denying the identity of Perez), the coder must know that Carlos Perez identifies as indígena and that he changed his first name to *Yaku*, a Quechua word. It is even more complicated for the machine-learning algorithm to make these connections solely from the information provided (i.e., the training set).

Fourth, for corpora covering more extended periods, researchers might also need to evaluate the robustness of the process to drift in discourse over time. While the different forms of racism are slow to change, their linguistic manifestations might not be. Furthermore, different sources of the text can also include different expressions. For example, the language used in social media will differ from that used in campaign ads or parliamentary speeches. If researchers find significant shifts in time and context, they must account for these in their training sets and validate their model across time and context.

Finally, our approach relies heavily on computational power. In Appendix H, we include a detailed explanation of the resources and platforms we used to run our models, together with a summary of their time and cost. We understand that not all researchers have access to the required computational power or the computational skills to implement a transformer-based machine learning approach. This perpetuates and accentuates the existing inequalities within the field and that it should not be an entry barrier to using state-of-the-art techniques. While we have tried to be as clear and comprehensible as possible, many elements of our process may be difficult to implement. Our future work also aims to provide a simple and efficient library that makes using a Transformers infrastructure accessible to everybody and applicable to various types of text.

[31] The original tweet reads: “Hasta hoy fui educada con este mestizo embaucador, andate a la verga Carlos Perez, eres cómplice del daño que le hacen a mi Ecuador reflechucha de tu madre.”

7 Conclusion

Systematically identifying racist discourse in large corpora has been a complex task (Van Dijk, 2005). Racism takes different forms, such as overt and covert racism, which are largely dependent on context. This article provides a methodological approach to classify different forms of racist discourse in text that combines a theoretical understanding of the context with learning approaches. Given the contextual nature of racism, we highlight the importance of creating coding rules that consider the origin and manifestations of racism in the place and time of study. Having theoretically grounded rules to categorize different forms of discursive racism is the first step to identifying them in large corpora. To this end, we provide a step-by-step approach to building a coding scheme to label different forms of racist discourse.

The second part uses a Transformers-based deep-learning approach to text classification. The Transformers architecture has revolutionized NLP tasks, partly for its ability to better understand context compared to other deep learning approaches, like CNN and RNN models. Transformers-based models such as XLM-RoBERTa can be further pre-trained with specialized text that improves classification performance. Adequately trained models can accurately classify text that requires a nuanced understanding of context, such as racist discourse. We apply this process to identify covert and overt racism in a corpus of 2M+ tweets relating to the indígena community in Ecuador between 2018 and 2021. Our main model XLM-R-Racismo, an XLM-RoBERTa model further pre-trained using 1.7 million tweets, outperforms other machine-learning models in detecting overt and covert racism in text.

The results from our analysis highlight not only the usefulness of our method but also the implications on how we understand the role of racist content in modern settings. We found the expected patterns of overtly racist content in social media: more central nodes (i.e., more important users) are less likely than more peripheral nodes to tweet racist messages. The effect was weaker on covertly racist tweets. The consequences of the characteristics of racist language in current-day Ecuador (i.e., language that is socially punished) are replicated in social media.

Finally, while the primary goal of this paper is to provide a step-by-step approach to classifying racist discourse in large corpora, we believe our contribution can be extended to other areas that require identifying coded language. Our approach focuses on a socially constructed notion such as race, and provides a nuanced identification of its various manifestations. There are other social constructions that can be identified following a similar process as the one described in this paper (e.g., benevolent vs. hostile sexism). Furthermore, it can be applied to other cases that employ coded language to communicate positions (and lower the costs of doing so). Dog-whistling, which has been employed to hide racist meaning–albeit not exclusively–, also requires contextual understanding and knowledge to be identified: the use of 'dog-whistling' has been reported in the United States (López, 2013), Sweden (Åkerlund, 2022), and Brazil (Sussi de Oliveira and Figuereo-Benítex, 2022), but is not yet salient in Ecuador.

The ability of Transformers-based approaches to understanding language in context makes them an ideal tool to classify text that requires a contextual understanding. Indeed, their effectiveness in annotation tasks, the larger tokenizer, and the potential for further training make these models especially well-suited for hate speech detection when compared to available alternatives. The identification of covert and overt racism in a highly specialized task serves as an example of the possibilities. Researchers can exploit the approach presented here to answer other substantive questions in the discipline, using the vast corpora available in new and creative ways.

References

Ahmed, Sara. 2012. On being included: Racism and diversity in institutional life. In *On being included.* Duke University Press.

Ahn, Hyeseon, Youngwook Kim, Jungin Kim and Yo-Sub Han. 2024. SharedCon: Implicit hate speech detection using shared semantics. In *Findings of the Association for Computational Linguistics ACL 2024*. pp. 10444–10455.

Åkerlund, Mathilda. 2022. "Dog whistling far-right code words: the case of 'culture enricher'on the Swedish web." *Information, Communication & Society* 25(12):1808–1825.

Albertson, Bethany L. 2015. "Dog-whistle politics: Multivocal communication and religious appeals." *Political Behavior* 37:3–26.

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.

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* pp. 102–125.

Blackington, Courtney and Frances Cayton. 2025. “To Dog-Whistle or to Bark? Elite Communication Strategies When Invoking Conspiracy Theories.” *Government and Opposition* 60(2):382–403.

Blinder, Scott, Robert Ford and Elisabeth Ivarsflaten. 2013. “The better angels of our nature: How the antiprejudice norm affects policy and party preferences in Great Britain and Germany.” *American Journal of Political Science* 57(4):841–857.

Bobo, Lawrence, James R Kluegel and Ryan A Smith. 1997. Laissez-faire racism: The crystallization of a kinder, gentler, antiblack ideology. In *Racial attitudes in the 1990s: Continuity and change*, ed. Jack Martin and Steven A Tuch. Praeger Westport, CT pp. 15–42.

Bonfil Batalla, Guillermo. 1977. “El concepto de indio en América: una categoría de la situación colonial.” *Boletín Bibliográfico de Antropología Americana (1973-1979)* 39(48):17–32.

Bonilla-Silva, Eduardo. 2001. *White supremacy and racism in the post-civil rights era.* Lynne Rienner Publishers.

Bonilla-Silva, Eduardo. 2002. “The linguistics of color blind racism: How to talk nasty about blacks without sounding “racist”.” *Critical Sociology* 28(1-2):41–64.

Bonilla-Silva, Eduardo. 2006. *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States.* Rowman & Littlefield Publishers.

Bonilla-Silva, Eduardo. 2013. “New racism,” color-blind racism, and the future of Whiteness in America. In *White out.* Routledge pp. 268–281.

Bonilla-Silva, Eduardo. 2015. “The structure of racism in color-blind,“post-racial” America.” *American Behavioral Scientist* 59(11):1358–1376.

Bridges, Khiara M. 2019. *Critical race theory: A primer.* Foundation Press.

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

Canessa, Andrew. 2007. “Who is indigenous? Self-identification, indigeneity, and claims to justice in contemporary Bolivia.” *Urban Anthropology and Studies of Cultural Systems and World Economic Development* pp. 195–237.

Chin, William Y. 2015. “The age of covert racism in the era of the roberts court during the waning of affirmative action.” *Rutgers Race & L. Rev.* 16:1.

Coates, Rodney D. 2008. “Covert Racism in the USA and Globally.” *Sociology Compass* 2(1):208–231.

Coates, Rodney D. 2011. Covert racism: An Introduction. In *Covert Racism*, ed. Rodney D Coates. Brill pp. 1–16.

Colloredo-Mansfeld, Rudi. 1998. “‘Dirty Indians’, radical indígenas, and the political economy of social difference in modern Ecuador.” *Bulletin of Latin American Research* 17(2):185–205.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer and Veselin Stoyanov. 2019. "Unsupervised cross-lingual representation learning at scale." *arXiv preprint arXiv:1911.02116* .

De la Torre, Carlos. 1996. *El racismo en Ecuador: experiencias de los indios de clase media.* Centro Andino de Acción Popular-CAAP.

Elias, Amanuel. 2024. "Racism as neglect and denial." *Ethnic and Racial Studies* 47(3):483–505.

ElSherief, Mai, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury and Diyi Yang. 2021. "Latent hatred: A benchmark for understanding implicit hate speech." *arXiv preprint arXiv:2109.05322* .

Essed, Philomena. 1991. *Understanding everyday racism: An interdisciplinary theory.* Vol. 2 Sage.

Gagliardone, Iginio, Matti Pohjonen, Zenebe Beyene, Abdissa Zerai, Gerawork Aynekulu, Mesfin Bekalu, Jonathan Bright, Mulatu Alemayehu Moges, Michael Seifu, Nicole Stremlau et al. 2016. "Mechachal: Online debates and elections in Ethiopia-from hate speech to engagement in social media." *Available at SSRN 2831369* .

Gambäck, Björn and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-Speech. In *Proceedings of the First Workshop on Abusive Language Online*, ed. Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy and Joel Tetreault. Vancouver, BC, Canada: Association for Computational Linguistics pp. 85–90.
URL: *https://aclanthology.org/W17-3013/*

Gao, Lei, Alexis Kuppersmith and Ruihong Huang. 2017. "Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach." *arXiv preprint arXiv:1710.07394* .

Gertner, Abigail S et al. 2019. MITRE at SemEval-2019 task 5: Transfer learning for multilingual hate speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation.* pp. 453–459.

Gilardi, Fabrizio, Meysam Alizadeh and Maël Kubli. 2023. "ChatGPT outperforms crowd workers for text-annotation tasks." *Proceedings of the National Academy of Sciences* 120(30):e2305016120.

Glover, Karen S. 2009. *Racial profiling: Research, racism, and resistance.* Rowman & Littlefield Publishers.

Grimmer, Justin, Margaret E Roberts and Brandon M Stewart. 2022. *Text as data: A new framework for machine learning and the social sciences.* Princeton University Press.

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):555–590.

Hampton, Lynn A. 2010. "Covert racism and the formation of social capital among a volunteer youth corps." *Critical Sociology* 36(2):285–305.

Henry, Patrick J and David O Sears. 2002. "The symbolic racism 2000 scale." *Political psychology* 23(2):253–283.

Herzog, Benno and Arturo Lance Porfillio. 2022. "Talking with racists: insights from discourse and communication studies on the containment of far-right movements." *Humanities and Social Sciences Communications* 9(1):1–7.

Hill, Jane H. 1995. "Junk Spanish, covert racism, and the (leaky) boundary between public and private spheres." *Pragmatics* 5(2):197–212.

Holdaway, Simon and Megan O'neill. 2007. "Where has all the racism gone? Views of racism within constabularies after Macpherson." *Ethnic and racial Studies* 30(3):397–415.

Huang, Fan, Haewoon Kwak and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*. pp. 294–297.

Katzew, Ilona. 2005. *Casta painting: images of race in eighteenth-century Mexico*. Yale University Press.

Khodak, Mikhail, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart and Sanjeev Arora. 2018. "A la carte embedding: Cheap but effective induction of semantic feature vectors." *arXiv preprint arXiv:1805.05388* .

Kim, Youngwook, Shinwoo Park and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th international conference on computational linguistics*. pp. 6667–6679.

Kim, Youngwook, Shinwoo Park, Youngsoo Namgoong and Yo-Sub Han. 2023. ConPrompt: Pre-training a language model with machine-generated data for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. pp. 10964–10980.

Kinder, Donald R and David O Sears. 1981. "Prejudice and politics: Symbolic racism versus racial threats to the good life." *Journal of personality and social psychology* 40(3):414.

Kinder, Donald R and Lynn M Sanders. 1996. *Divided by color: Racial politics and democratic ideals*. University of Chicago Press.

Kroskrity, Paul V. 2021. "Covert linguistic racisms and the (re-) production of white supremacy." *Journal of Linguistic Anthropology* 31(2):180–193.

Kruk, Julia, Michela Marchini, Rijul Magu, Caleb Ziems, David Muchlinski and Diyi Yang. 2024*a*. "Silent Signals, Loud Impact: LLMs for Word-Sense Disambiguation of Coded Dog Whistles.".
URL: *https://arxiv.org/abs/2406.06840*

Kruk, Julia, Michela Marchini, Rijul Magu, Caleb Ziems, David Muchlinski and Diyi Yang. 2024*b*. "Silent Signals, Loud Impact: LLMs for Word-Sense Disambiguation of Coded Dog Whistles." *arXiv preprint arXiv:2406.06840* .

Levchak, Charisse C. 2018. Microaggressions, Macroaggressions, and Modern Racism. In *Microaggressions and modern racism*. Springer pp. 13–69.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. 2019. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* .

Loper, Edward and Steven Bird. 2002. "Nltk: The natural language toolkit." *arXiv cs/0205028* .

López, Ian Haney. 2013. *Dog whistle politics: How coded racial appeals have reinvented racism and wrecked the middle class*. Oxford University Press.

Loshchilov, Ilya and Frank Hutter. 2017. "Decoupled weight decay regularization." *arXiv:1711.05101* .

Lynch, Emily E, Lorraine Halinka Malcoe, Sarah E Laurent, Jason Richardson, Bruce C Mitchell and Helen CS Meier. 2021. "The legacy of structural racism: associations between historic redlining, current mortgage lending, and health." *SSM-population health* 14:100793.

Marom, Lilach. 2019. "Under the cloak of professionalism: Covert racism in teacher education." *Race Ethnicity and Education* 22(3):319–337.

Martínez Novo, Carmen. 2018. "Ventriloquism, racism and the politics of decoloniality in Ecuador." *Cultural studies* 32(3):389–413.

Murrey, Amber and Nicholas A Jackson. 2020. "A decolonial critique of the racialized "localwashing"' of extraction in Central Africa." *Annals of the American Association of Geographers* 110(3):917–940.

Oberem, Udo. 1985. "La sociedad indígena durante el Periodo Colonial de Hispanoamérica." *Miscelánea Antropológica Ecuatoriana* 5:161–218.

Ocampo, Nicolás Benjamín, Ekaterina Sviridova, Elena Cabrio and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *EACL 2023-17th Conference of the European Chapter of the Association for Computational Linguistics*. Vol. 2023 Association for Computational Linguistics pp. 1997–2013.

Ousidhoum, Nedjma, Zizheng Lin, Hongming Zhang, Yangqiu Song and Dit-Yan Yeung. 2019. Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, ed. Kentaro Inui, Jing Jiang, Vincent Ng and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics pp. 4675–4684.
URL: *https://aclanthology.org/D19-1474/*

Painter, Nell Irvin. 2010. *The history of white people*. WW Norton & Company.

Pallares, Amalia. 2002. *From peasant struggles to Indian resistance: The Ecuadorian Andes in the late twentieth century*. University of Oklahoma Press.

Palmer, Alexis, Noah A Smith and Arthur Spirling. 2024. "Using proprietary language models in academic research requires explicit justification." *Nature Computational Science* 4(1):2–3.

Pierson, Emma, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff et al. 2020. "A large-scale analysis of racial disparities in police stops across the United States." *Nature human behaviour* 4(7):736–745.

Plaza-del Arco, Flor Miriam, M Dolores Molina-González, L Alfonso Ureña-López and M Teresa Martín-Valdivia. 2021. "Comparing pre-trained language models for Spanish hate speech detection." *Expert Systems with Applications* 166:114120.

Pons, Pascal and Matthieu Latapy. 2005. Computing communities in large networks using random walks. In *International symposium on computer and information sciences*. Springer pp. 284–293.

Rheault, Ludovic and Christopher Cochrane. 2020. "Word embeddings for the analysis of ideological placement in parliamentary corpora." *Political Analysis* 28(1):112–133.

Rich, Paul B et al. 1990. *Race and empire in British politics*. CUP Archive.

Rodriguez, Pedro L. and Arthur Spirling. 2022. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." *The Journal of Politics* 0(0):000–000.

Roitman, Karem and Alexis Oviedo. 2017. "Mestizo racism in Ecuador." *Ethnic and racial studies* 40(15):2768–2786.

Saha, Koustuv, Eshwar Chandrasekharan and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM conference on web science.* pp. 255–264.

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.* pp. 1–10.

Schreier, Margrit. 2012. *Qualitative content analysis in practice.* Sage publications.

Shoshana, Avihu. 2016. "The language of everyday racism and microaggression in the workplace: Palestinian professionals in Israel." *Ethnic and Racial Studies* 39(6):1052–1069.

Sussi de Oliveira, Juliete and Juan C Figuereo-Benítex. 2022. "'Dog Whistle' en los discursos de Jair Bolsonaro y Santiago Abascal a través de YouTube." *Contenidos, medios e imágenes en la comunicación política* .

Telles, Edward and Stanley Bailey. 2013. "Understanding Latin American beliefs about racial inequality." *American Journal of Sociology* 118(6):1559–1595.

Timoneda, Joan C. and Sebastián Vallejo Vera. 2025*a*. "BERT, RoBERTa or DeBERTa? Comparing Performance Across Transformer Models in Political Science Text." *The Journal of Politics* 00(00):00.

Timoneda, Joan C and Sebastián Vallejo Vera. 2025*b*. "Memory Is All You Need: Testing How Model Memory Affects LLM Performance in Annotation Tasks." *arXiv preprint arXiv:2503.04874* .

Traverso-Yépez, Martha. 2005. "Discursos racistas: institucionalización del racismo a través de las prácticas lingüísticas." *Revista Interamericana de Psicología* 39(1):61–70.

Tunstall, Lewis, Leandro von Werra and Thomas Wolf. 2022. *Natural Language Processing with Transformers.* O'Reilly.

Uyheng, Joshua, Daniele Bellutta and Kathleen M Carley. 2022. "Bots amplify and redirect hate speech in online discourse about racism during the COVID-19 pandemic." *Social Media+ Society* 8(3):20563051221104749.

Vallejo Vera, Sebastián. 2023. "Rage in the Machine: Activation of Racist Content in Social Media." *Latin American Politics and Society* 65(1):74–100.

Van Dijk, Teun A. 1993. *Elite discourse and racism.* Sage Pulbications, Inc.

Van Dijk, Teun A. 2005. *Racism and discourse in Spain and Latin America.* John Benjamins Publishing Company Amsterdam.

Van Dijk, Teun A. 2009. Racism and discourse in Latin America: An introduction. In *Racism and discourse in Latin America*, ed. Teun A Van Dijk. Rowman & Littlefield Publishers pp. 4–13.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems.* pp. 5998–6008.

Walsh, Keith R. 2004. "Color-blind racism in Grutter and Gratz." *Boston College Third World Law Journal* 24(2):443.

Whitten Jr, Norman E. 2003. “Symbolic inversion, the topology of El Mestizaje, and the spaces of Las Razasin Ecuador.” *Journal of Latin American Anthropology* 8(1):52–85.

Wright, Michelle. 2004. *Becoming black: Creating identity in the African diaspora.* Duke U. Press.

Wu, Tianhao, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston and Sainbayar Sukhbaatar. 2024. “Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge.” *arXiv preprint arXiv:2407.19594* .

Xu, Wen. 2025. “Linguistic racism and micro-aggressions in everyday encounters of African migrants in China: A challenge to the nation’s strategic vision for Africa?” *Ethnicities* 25(1):69–84.

Yang, Kai-Cheng, Onur Varol, Pik-Mai Hui and Filippo Menczer. 2020. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence.* Vol. 34 pp. 1096–1103.

Yoder, Michael, Lynnette Ng, David West Brown and Kathleen Carley. 2022. How Hate Speech Varies by Target Identity: A Computational Analysis. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, ed. Antske Fokkens and Vivek Srikumar. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics pp. 27–39.
URL: *https://aclanthology.org/2022.conll-1.3/*