


# Toxic Social Media: Affective Polarization After Feminist Protests

Marcela Suarez Estrada<sup>1</sup> , Yulissa Juarez<sup>2</sup>,  
and C. A. Piña-García<sup>2</sup>

Social Media + Society  
April-June 2022: 1–12  
© The Author(s) 2022  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/20563051221098343  
journals.sagepub.com/home/sms  


## Abstract

The objective of this article is to conceptualize affective polarization beyond partisan politics to instead analyze the ways in which women's affective political participation is subject to toxic discipline. While a lot of focus has been placed on affective politics as mechanisms for governance, little has been done regarding affective polarization after feminist protest. In this article, we bridge two bodies of literature— affective politics and political polarization— by proposing the notion of affective polarization. We focused on the case of a series of feminist mobilizations that took place to fight back against the impunity of police violence in Mexico. We conducted a mixed-method approach that combines, on one hand, quantitative analysis of data strand tweets encompassing #EllasNoMeRepresentan (TheyDoNotRepresentMe) ( $N = 17,698$ ) and #EllasSiMeRepresentan (TheyDoRepresentMe) ( $N = 6700$ ) and, on the other hand, a qualitative analysis of 500 tweets of each hashtag. The results of the study revealed the existence of polarization that aims at disciplining the affective political participation of women. Almost half of our data contain negative sentiments. The toxic tweets include corrective threats, such as incitation to sexual violence, murder, hate against feminism, and patronizing discourses about how women should protest. We thus conclude that while it is true that social media has amplified feminist mobilization, it has also led to an increase of digital violence. With these findings, the article contributes to a better understanding of both feminist affective politics and its disciplining governing mechanisms in a patriarchal social media.

## Keywords

digital violence, toxic social media, sentiment analysis, feminist mobilization, affective politics

## Introduction

The serious situation of gender violence affecting Mexico has caused intense feminist protests over the last 5 years. These mobilizations have shown new ways of doing affective feminist politics. Women are mobilizing new forms of political repertoires subverting the very affects that have thus far been used to govern their bodies: hate and guilt (Penz & Sauer, 2020). By showing themselves to be fearless and full of rage in the political arena, for example, or by intervening at historical monuments with graffiti. Furthermore, they aim to hold the state accountable for the various ways that they are affected by violence.

While there is an extensive literature on feminist politics appropriating social media to deploy tactics for political mobilization, such as hashtags (Davis & Santillana, 2019), the literature tends to remain somewhat triumphalist, lauding the new tools for mobilization. The social media situation for feminist politics is all but great; however, in Mexico,

increasing digital violence is targeting feminist protesters (Amnistía Internacional, 2021). This violence is part of coordinated attacks, threats of death and rape via explicitly violent imagery. While this issue is getting increasing attention in the literature (Ananías Soto & Sánchez Vergara, 2019; Gallacher et al., 2021), we maintain that there is a need for more studies approaching this issue that puts the focus on the affects—not only as driver of social mobilizations but also as a site for disciplining the ways in which women participate in political life. The article asks the following research question: In which ways are the affects of

<sup>1</sup>Freie Universität Berlin, Germany

<sup>2</sup>Universidad Veracruzana, Mexico

### Corresponding Author:

Marcela Suarez Estrada, Lateinamerika-Institut, Freie Universität Berlin, Rüdeshheimer Str. 54-56, 14197 Berlin, Germany.

Email: marcela.suarez@fu-berlin.de

Twitter: @marentierras



feminist protesters subject to toxic discipline and affective polarization in social media? It contributes to a better understanding of both feminist affective politics and the mechanisms by which it gets disciplined, such as polarization and toxicity in our digital society.

In this article, we want to fill this gap through a mixed (quantitative and qualitative) methodological approach that entails the analysis of a corpus of tweets that emerged from feminist mobilizations with the hashtags #TheyDoNotRepresentMe and #TheyDoRepresentMe. The results of our study show that, in the wake of the protests, women were subject to polarization and digital violence on social media platforms. We contend that the affective discipline after feminist mobilizations deserves more attention to render visible polarization and toxicity as gender disciplining strategies. It is important to note that our collected tweets need to be interpreted with caution since this dataset includes strong language, such as: depictions of violence, misogynistic, or other discriminatory language or behavior.

### **Affective Politics in Social Media**

Affect theory has put the focus on the diverse ways in which humans are governed through affects (Jupp et al., 2016). Feminist theory has contributed with discussions about materiality in the governing practices that control women bodies through shame and fear (Penz & Sauer, 2020; Ward et al., 2019) while being encoded in cultural patterns (Ahmed, 2014; Savigny, 2020). Affects are thus political because they shape collective social order. For instance, they reinforce binary gender orders that construct women as emotional and weak, and men as rational and emotionless (Bargetz, 2015; Penz & Sauer, 2020). Politically active women become the target of attacks as a way of excluding them from political life (Åhäll, 2018; Krook, 2020b).

The recent contributions on affect theory have also discussed the ways in which affects are intertwined with media practices (Hynnä et al., 2019; Papacharissi, 2016). These contributions have brought to the front feminist discussions of embodiment, materiality, and power to contest the idea of the apparently neutral and objective space (Sundén & Paasonen, 2020; Zarzycka & Olivieri, 2017). Technologies are forming assemblages not only made of technologies but also affects (Bennett, 2010; Deleuze & Guattari, 1987; Hiilis et al., 2019). Moreover, the affective governance practices are not limited to the technologies of governing the self; they are also translated into profit on social media (Hynnä et al., 2019). For that reason, political disagreement has become highly affective and thus subject to platform governing mechanisms.

Social media is also a space where the circulation of gender stereotypes is mobilized and questioned. This is made through the materialization of power discourses in text,

images, memes, and more diverse content. These gender discourses work affectively, emotionally, and performatively through shared meanings, perceptions, and social norms (Åhäll, 2018, p. 43). Since affect is not only a site of oppression but also a site of political transformation (Pedwell & Whitehead, 2012), feminist affective politics literature have disputed such gender stereotypes and the dichotomies between rationality and affectivity. This literature has also discussed the role of affects (hope, solidarity, and rage) in the political arena for feminist purposes (Hemmings, 2012; Savigny, 2020).

Affective practices in social media mobilize affects understood as cultural products governed by norms of what and how we should feel, express, and do emotions (Döveling et al., 2018). In this way in hashtags circulate collective affective practices of shared perceptions (Ahmed, 2014). These practices shape collective regimes to discipline affects (Reckwitz, 2016). Although we claim that these governing practices are not new, we think that their governing has been socially escalated in social media through various strategies. One of these strategies is polarization. Through the division into “us” (non-feminists) and “them” (the crazy feminists), an affective anti-feminism discourse is created that mobilizes hate, shame, and fear. In that discourse, affective regimes arise about how citizens should feel after feminist mobilizations. Another strategy is the diffusion of toxic speech with incitation to sexual violence and murder.

This study contributes to affective politics literature by shifting the focus away from a triumphalist vision on protests and feminist mobilization to rather center on the polarization and toxic speech that arises after the protests. While a lot focus has been placed on affective polarization in partisan politics (Kelly Garrett et al., 2019), little attention has been paid to polarization strategies to discipline women’s affects and thus their participation in politics. In this article, we shed light on this research gap.

### **Feminist Political Mobilizations in Mexico**

In this section, we will focus on the case of two feminist protests that were subject to digital violence. In August 2019 in a 2-week period of time, two women were raped at the hands of police officers in Mexico City. These cases are not isolated cases of the systemic violence and feminicides that have increased significantly from 1995 on. There are several obstacles women face when they come forward to denounce the perpetrators, such as being blamed and shamed based on gender stereotypes (Arjona Estévez, 2019), for instance, for drinking alcohol or partying at the moment the crime occurred. In these two cases, series of irregularities occurred in the collection of samples by forensic experts and one of the victim’s personal data were leaked by authorities. Leaking a victim’s data especially

gives evidence of how the fear of revictimization is used to govern women in Mexico.

In this context, feminist mobilizations have included the new affective repertoires based on the subversion of fear and shame. This subversion is exemplified in one the feminist protest cases analyzed here that took place after the minor rape cases in August 2019 in Mexico City, known as *Brillantada* (the *Glitter Protest*). The name *Brillantada* is due to the fact that the protesters dumped glitter on the head of Mexico City's Security Secretary in the middle of an interview. After that, the Chief of Government of Mexico City said "this was not a protest; it was a provocation." Full of indignation for the response of the female mayor of Mexico City, feminist collectives called for a mobilization to vindicate their right to protest. The protests took place on August 16, 2019. It started in front of bus station in front of the Secretaría de Seguridad Pública. Women smashed windows with advertising messages that depicted women as aesthetic models. They also spray-painted slogans that read "not one more." The protest continued to the historical monuments, the Hemiciclo a Juárez and the Ángel de la Independencia. On the way to the main historical monuments, protesters passed by a police station. They replicated the same practices: painted feminist slogans, such as "Estado Femicida" (Femicide State).

The day after the protest, a big discussion arose on Twitter about how violent the women had been. Also, it was claimed that with that kind of protest they would not achieve anything. Some tweets, through memes, referred to female religious figures, such as Sor Juana Inés de la Cruz,<sup>1</sup> as examples of how women should provoke social change: peacefully and with class. Thus, evoking an alienation of affections (Döveling et al., 2018) about how to feel after the feminist protest and thus reinforcing affective gender stereotypes (Åhäll, 2018; Ahmed, 2014). These accusations were deployed through the hashtags #TheyDoNotRepresentMe which began to trend on Twitter very quickly. Just 2 hr after, the #TheyDoRepresentMe was created. The meaning of the first hashtag renders visible the creation of a collective enemy of "them"—women who are violent—and "us" who are against these forms of protest. These framed the public discussion around the ways in which women were led by their affects based on rage to "destroy" historical monuments. Meanwhile, the second hashtag meant to vindicate the civil right to protest and discuss the affectations of violence against women.

The other huge feminist protest that followed 1 year after the *Brillantada* in August 2019 was the 8 March protest in 2020. Prior to this second case, an affective atmosphere plagued by fear had prevailed in Mexico City in the interim. On one hand, collective threats began to circulate on social networks, specifically to throw acid on women who dare attend the feminist protest scheduled for International

Women's Day. The atmosphere online was so threatening, even the government issued a statement saying that it would investigate these threats. Although the cyber police tried to appease the fear, so that, women would assert their right to protest (Ruiz, 2020), the government mobilized, on the other hand, a large number of police and barriers to cover the main historical monuments and government buildings, based on fear that women would graffiti them. The same tweets that were created in August 2019 were activated again. The press turned the focus back to women graffitiing the barriers surrounding the historical monuments and not in the police violence. The hashtag #TheyDoNotRepresentMe was reactivated to continue the discussion of how violent the women protesters were being, as well as #TheyDoRepresentMe to defend the cause.

## Methods

This study approaches through a mixed methods framework comprised quantitative and qualitative analysis of tweets published after feminist protests in the period from 16 August 2019 to 20 March 2020. We selected this period as it covers the *Brillantada* and the "International Women's Day" (8 March, 2020) protest. The quantitative approach entails a sentiment, toxicity, and cluster semantic analysis. While the qualitative entails an analysis of 500 tweets collected on our datasets.

We have collected publicly available tweets from 16 August 2019 to 20 March 2020 (time window) via the Twitter streaming application programming interface (API). According to the privacy policy, this research inspected only those tweets that were public (i.e., no privacy settings were selected by the user). With the aim to comply with Twitter's terms of service, data cannot be publicly shared. Interested future researchers may reproduce the experiments by following the procedure described in the following part of this article. We also include two additional measures to ensure the privacy of participants: data anonymization and changes in the content of the tweets to avoid de-anonymization.

Using a customized query string to filter tweets, we have gathered specific tweets that contain at least one of the following Spanish language hashtags: #EllasNoMeRepresentan and #EllasSiMeRepresentan.

We developed a "social explorer" to retrieve data from Twitter via its standard API. To replicate this study, we highly recommend to go along with the official step-by-step guide provided on the following link: <https://developer.twitter.com/en/docs/tutorials/step-by-step-guide-to-making-your-first-request-to-the-twitter-api-v2> this guide describes how to make a request using different coding languages, such as: Java, Node.js, Python, R, and Ruby. A python script of our Twitter social explorer is showed as follows:

```

#This is where you initialize the client with your own bearer token (replace the XXXXXXXX
with your own bearer token)
client = Twarc2(bearer_token="XXXXXXXXXX")

# Specify the start time in UTC for the time period you want Tweets from
start_time = datetime.datetime(2019, 8, 16, 0, 0, 0, datetime.timezone.utc)

# Specify the end time in UTC for the time period you want Tweets from
end_time = datetime.datetime(2020, 3, 20, 0, 0, 0, datetime.timezone.utc)

#we specify our queries
query = "#EllasNoMeRepresentan OR #EllasSiMeRepresentan" -is:retweet lang:es"

# The counts_all method call the full-archive Tweet counts endpoint to get Tweet volume
based on the query
count_results = client.counts_all(query=query, start_time=start_time, end_time=end_time)

# The search_all method call the full-archive search endpoint to get Tweets based on the
query, start and end times
search_results = client.search_all(query=query, start_time=start_time, end_time=end_time,
max_results=10)

```

Our sample consisted of 17,698 tweets related to #TheyDoNotRepresentMe and 6,700 tweets related to #TheyDoRepresentMe. These datasets contain information, such as: user ID, the screen name or alias, number of followers, date, text, device used to post the tweet (source), and the user-defined location. To detect, filter, and remove corrupt or inaccurate tweets; we carried out a process of data cleansing and data management. To begin this process, we removed errors, such as nulled fields, empty sets, and incomplete data. After collection of our dataset, we discarded off-topic tweets in a semi-automated way by filtering only those tweets that were posted repeatedly. Our filtered dataset is as follows  $N=14,035$  for #TheyDoRepresentMe and  $N=5,245$  for #TheyDoNotRepresentMe.

In this research, we have considered the morphological characteristics of language, such as when text is broken into sequences of characters. Thus, the idea to conduct our text analysis was through representing raw text as numbers with the aim to perform computation on them. Essentially, natural language must be transformed to a machine-readable, numeric representation to be ready for computation. This is what makes our study both positivist while incorporating a critical feminist reading of the text.

After the collection and filtering processes, our sample datasets were explored using machine learning models based on the perspective API (see: <https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages>). This API allows us to identify abusive tweets. Essentially, these models score a phrase based on the perceived impact the text may have in a conversation. The perspective API evaluates the input text across a range of emotional concepts, called attributes. This attribute provides an overall score for the whole comment in this case for each tweet, obtaining as a result a probability output with a value between 0 and 1.

Higher values indicate a higher probability that the tweet is toxic.

In this regard, the perspective API provides the following list of attributes: toxicity, severe\_toxicity, identity\_attack, insult, profanity, and threat. It is important to note that perspective's scores are an indication of probability, not severity, that is, higher numbers represent a higher likelihood that the patterns in the text resemble patterns in comments that people have tagged as toxic. The number is not a score of how toxic a particular entry is, but that it is likely to also be a toxic tweet. With these scores, we picked a threshold ( $\text{Thr} \geq 0.6$ ) to determine the likelihood of toxicity of a tweet.

It should be highlighted that we manually check every single tweet to determine this threshold, that is, we have combined the native machine learning models in the perspective API with a manual inspection with the aim to get better insights about the perceived impact a comment might have on a Twitter conversation. In addition, a higher score indicates a greater likelihood that a reader would perceive the comment as containing the given attribute (toxicity, severe\_toxicity, identity\_attack, insult, profanity, and threat).

Once the toxicity scores were computed, a sentiment analysis was carried out. Similar to the toxicity analysis, another API was used to assess polarity of the tweets. We picked the sentiment analysis API provided by MeaningCloud (<https://www.meaningcloud.com/products/sentiment-analysis>). This online tool is based on a semantic approach and advanced natural language features, such as: morphology, syntax, semantics, and pragmatics. Thus, this tool generates a syntactic-semantic tree of the text, and over this, terms of the lexicon are applied to spread their polarity values along the tree, properly combining the values depending on the morphological category of the word and the syntactic relations that affect them.

**Table 1.** Sentiment Analysis Results.

Hashtag	Number of tweets	Number of tweets clean	Sentiment analysis					No sentiment
			Negative	Very negative	Neutral	Positive	Very positive	
#TheyDoNotRepresentMe	17,698	14,035	4,689	1,818	1,449	3,196	649	2,234
Percentage		100%	33.4	13	10.3	22.8	4.6	15.9
#TheyDoRepresentMe	6,700	5,245	1,702	541	556	1,389	227	830
Percentage		100%	32.4	10.3	10.6	26.5	4.3	15.8

Source: Own elaboration.

In addition to the overall polarity of the text, the engine returns the polarity for word groups or segments of the text, in six possible levels: positive (P) and negative (N), very positive (P+) and very negative (N+), neutral (NEU), and none (NONE) in the event that no polarity is involved. Based on this, our sentiment analysis was carried out with the aim to determine whether a tweet expressed a positive/negative/neutral sentiment. Therefore, we were able to obtain a polarity of every tweet at record level. It should be noted that this API gives the user the possibility of detecting the polarity of user-defined entities and concepts, making the tool applicable and flexible to any kind of scenario. Thus, we customized our own dictionary and model with the aim to match those words with the Mexican context. Our dataset was structured according to the abovementioned six polarity levels. Then, we explore the datasets with a cluster hierarchical and semantic analysis to identify conversation clusters.

A hierarchical cluster analysis was performed to explore the semantic relationship of the words. This allows us to see what conversations were generated around each hashtag. The hierarchical cluster method was used to find similarities or differences between the observations of the groups. The Euclidean distance was applied in the tweets obtaining the similarity or difference between the words of the clusters. In terms of the semantic analysis, a tree diagram (dendrogram, Figure 3) using a hierarchical clustering algorithm was generated. First, the algorithm calculates the Euclidean distance, that is, it measures how similar are words to each other. Words included in the dendrogram are those that were most prevalent. Therefore, those words that are linked with each other show a closer hierarchical relationship.

From the qualitative analysis point of view, we have chosen three types of tweets: (1) The ones that made explicit influencing feelings or emotions, such as rage, hate, and violence (i.e., incitation to murder, rape), and also hope, empowerment, and solidarity. (2) The tweets that reinforcement of specific gender affective stereotypes, such as the feminist protesters are violent, crazy, whores, and wild beast. Similarly, we select those tweets that dispute these stereotypes mobilizing women being furious and politically active. Women who are not victims waiting for justice but women who perform their own ways of achieving political goals. (3) We focused on tweets that reinforce specific

affective dichotomies, such as the division of intelligence/affectivity and at the same time tweets that challenge these dichotomies.

Qualitative analysis was carried out in two stages. First, based on the result of the toxicity and sentiment analysis, we examined qualitatively the top 10 most toxic tweets in both hashtags. Then, the two more toxic tweets for each hashtag were assessed based on feminist literature on affect theory and affective feminist politics.

The second stage refers to qualitatively explore the results of our cluster analysis with the aim to identify and provide some context from the explored hashtags. Finally, we studied the relation between gender stereotypes, affective registers and dichotomies.

## Results

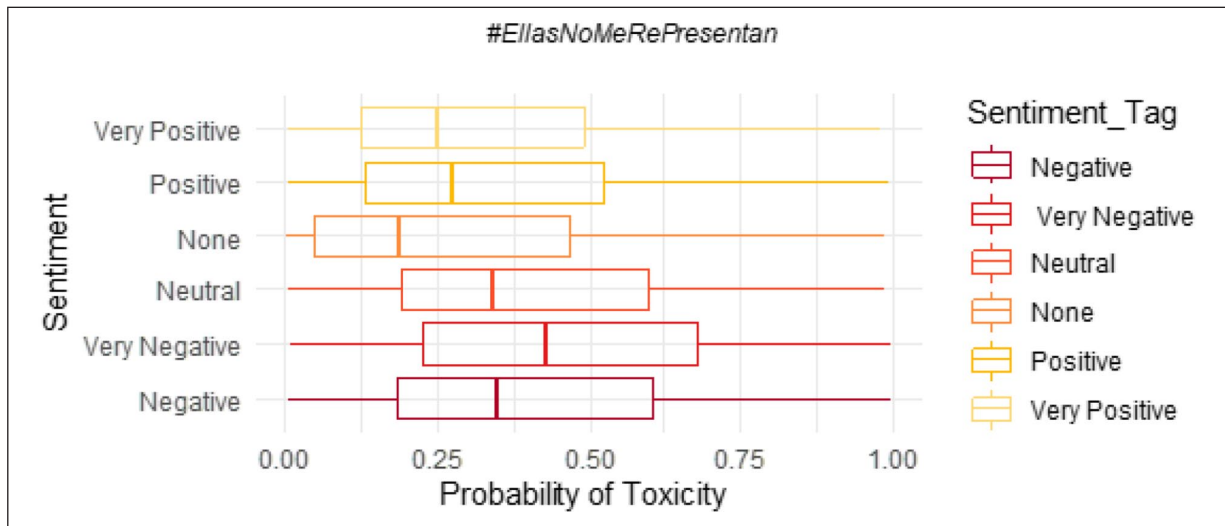
Table 1 provides preliminary results about the sentiment analysis; it is apparent from this table that 46.4% of the #TheyDoNotRepresentMe tweets had negative and very negative sentiment. On the other hand, tweets with positive and very positive sentiment were only 27.4%. It should be noted that tweets with neutral sentiment or no sentiment made up a quarter (25%) of the total number of tweets. #TheyDoRepresentMe hashtag had a similar trend. Negative and very negative sentiment predominated with 32.4% and 10.3% respectively, 43.7% of all the tweets in our dataset. In total, Table 1 shows that there is a polarization between both hashtags where negative sentiments were notable. However, it is necessary to highlight that the number of #TheyDoNotRepresentMe tweets is almost three times greater than #TheyDoRepresentMe and that many of the tweets against women protesters used #TheyDoRepresentMe in the same way. We carried out a toxicity assessment with the aim to get better insights from our collected tweets. Table 2 compares the results obtained from the preliminary analysis in terms of the measures of central tendency and dispersion of the hashtags.

Results from Table 2 reveal that tweets with negative sentiments were the most toxic in both hashtags. Although #TheyDoNotRepresentMe is slightly more toxic than #TheyDoRepresentMe, in both cases, the very negative tweets shown higher levels of toxicity. According to Table 2,

**Table 2.** Toxicity and Sentiment: Measures of Central Tendency and Dispersion of the Hashtags.

Sentiment	#TheyDoNotRepresentMe			#TheyDoRepresentMe		
	Mean	Median	Variance	Mean	Median	Variance
Very positive	0.330	0.249	0.0628	0.292	0.239	0.0561
Positive	0.350	0.277	0.0689	0.287	0.232	0.0546
Neutral	0.404	0.343	0.0677	0.346	0.295	0.0579
Negative	0.407	0.347	0.0716	0.356	0.301	0.0582
Very negative	0.458	0.429	0.0738	0.429	0.386	0.0622
None	0.277	0.190	0.0713	0.234	0.158	0.0489

Source: Own elaboration.



**Figure 1.** Toxicity of the tweets in relation to sentiment #TheyDoNotRepresentMe.

there was more variability of #TheyDoNotRepresentMe in comparison to #TheyDoRepresentMe. This suggests that the data were not clustered only at a certain range of toxicity but were more homogeneously dispersed.

In Figure 1, it can be observed that the negative sentiment was that of highest average toxicity with a probability close to 0.50. This contrasts with the very positive and positive tweets that have a similar average toxicity of 0.25. The y-axis of the box plot represents the different sentiments evaluated in the tweets. The x-axis evaluates the probability that a tweet is toxic.

Moreover, Figure 1 compares the distribution of data across five box plots, in this regard, it can be seen that all these plots showed an asymmetrical distribution, that is,

probabilities occur at irregular frequencies and the mean, median, and mode occur at different points. For instance, the median in most cases displays values below 0.50.

Similarly, Figure 2 depicts an asymmetrical distribution with most of the levels of toxicity above the median. Interestingly, those tweets labeled as “None” showed atypical data outside the whiskers. In general terms, all the categories reported a median value below to 0.50.

In the following paragraphs, we will proceed to analyze the tweets qualitatively. Figures 1 and 2 also show the affective polarization expressed in high levels of toxicity in negative and very negative sentiments. According to the results of our analysis, the following four tweets are the most toxic of our datasets:

#TheyDoNotRepresentMe	#TheyDoRepresentMe
fucking bitches look like they were fucked by their grandfather and live with their cocks inside them	if you are more concerned about a painted monument and not about women being raped and murdered for the simple fact of being a woman, let me tell you that you are a piece of shit, a vulgar piece of shit
damn bad born bitches they are the ones that should be killed	it doesn't bother them that they have painted the walls and broken the windows they are afraid that women are no longer their bitches and that they have realized that they don't need men for absolutely nothing and that they will beat their fucking rapist motherfuckers

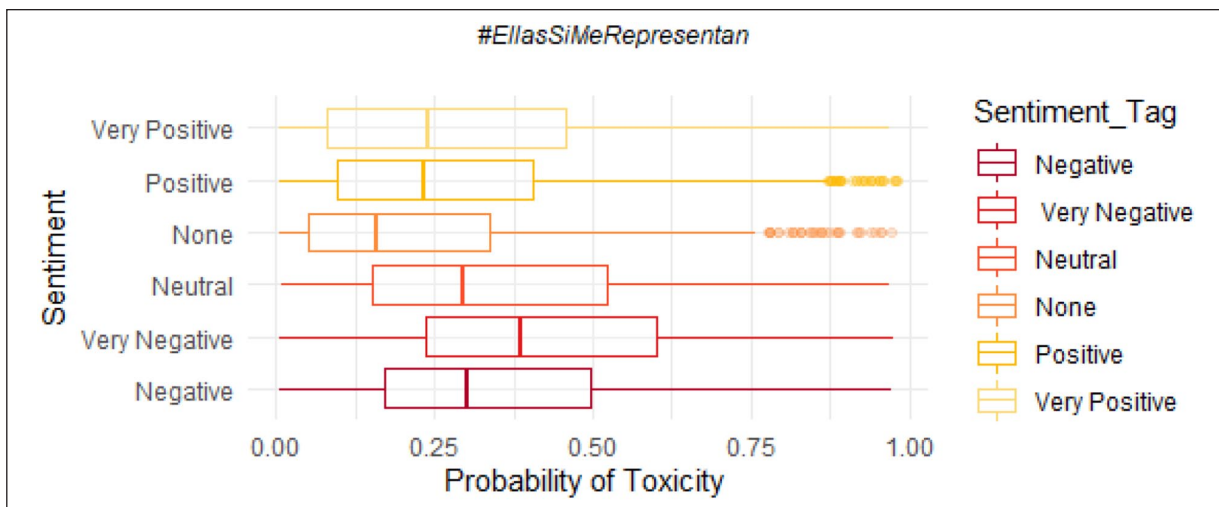


Figure 2. Toxicity of the tweets in relation to sentiment #TheyDoRepresentMe.

On one hand, the two tweets of #TheyDoNotRepresentMe exemplify the different extremely toxic political speeches directed at women protesters. The first mobilizes a discourse of intimidation, using the allegory of the sexual domination of men against women. In this case, represented in the grandfather as the patriarch of a family, who establishes his power over a woman through sexual domination. The reference to the phallus remaining in the female body can be interpreted to mean that such sexual dominance would continue to be perpetuated from generation to generation despite women’s undisciplined behavior. The second tweet is a clear incitement to femicide violence toward protesters. The threads of violence are strategies to exclude women from the political arena (Sanín, 2020). The #TheyDoRepresentMe set of tweets, on the other hand, denigratingly insults men who are more outraged by the graffitied monuments than by the murdered women. While the last tweet points out that men are afraid of women. The tweet ends the sentence with a threat to attack their rapist genealogy. These tweets are a sample of the toxic affectivity in both hashtags. However, the first hashtag is an expression of patriarchy and misogyny in social media, whereas the second one is a strategy to vindicate the political right to protest against state violence. It should be noted that some of the most toxic tweets of the hashtag #TheyDoRepresentMe had the same condemning tone as #TheyDoNotRepresentMe because users posted with both hashtags against women.

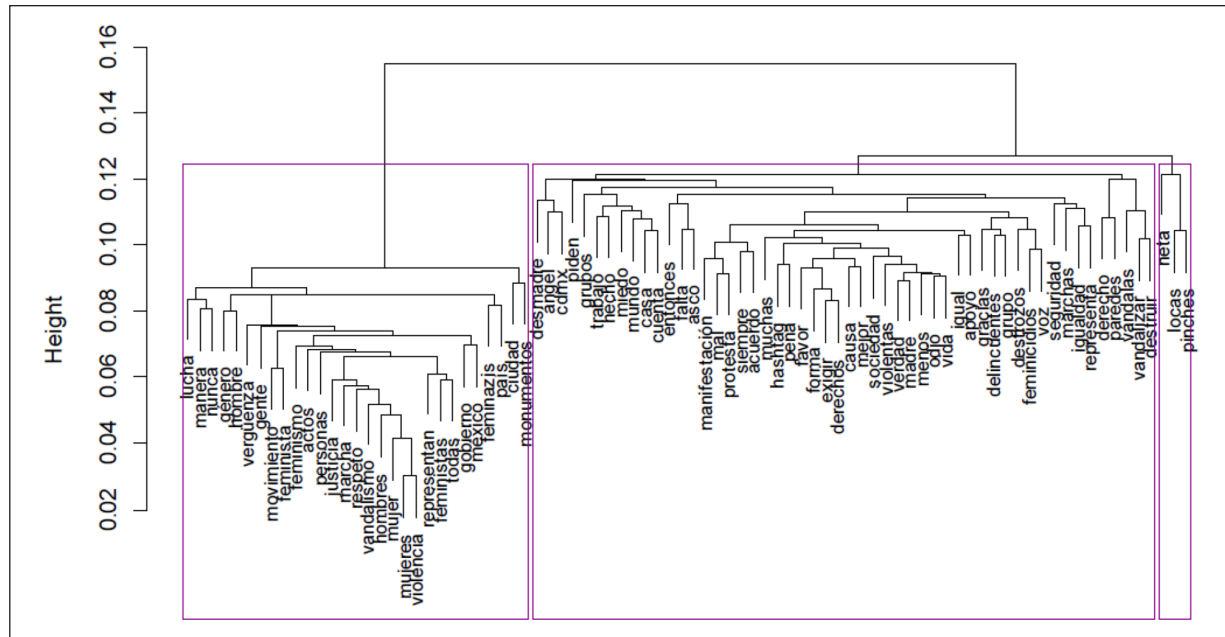
Figure 3 depicts the results of the hierarchical cluster analysis. There were three clusters formed.

The horizontal axis of Figure 3 stands for the cluster topics discussed and the number of clusters, while the vertical axis represents the how closely the terms topical clusters are related to one another. In the following paragraphs, we will analyze the content of the tweets that incorporated the keywords for each cluster. In the first cluster, the conversations were around topics, such as “lucha” (fight), “manera” (ways), “gender,” “mujeres (women),” “(violencia) violence,” “vergüenza (shame),” “justiciar (justice),” “gobierno (government),” “monumentos (monuments),” “feminazis,” and “ciudad (city).” During these Twitter conversations, several tweets showed discrimination against women for graffiti interventions at historical monuments. Others labeled the feminist interventions as violence and condemned them on the understanding that violence cannot be fought with more violence. Other tweets used the rhetorical strategy of exaggeration to point out that the feminist protests had destroyed the entire city. That is the reason why the word “city” appears in the dendrogram. A closer look into the tweets gives examples of the affective mechanisms that were mobilized, for example, in how the women should feel after the protest: ashamed. Feminist literature on affects has reported similar results about the way in which violence of women is communicated through feelings of pity, shock, and unease (Åhäll, 2018). The enunciation of this affect was followed by the use of denigrating words, such as feminazis, animals, bitches, or wild beasts. The following tweets are examples:

#TheyDon’tRepresentMe what a shame to see all the mess and violence provoked in a protest against violence, they look like animals instead of human beings, yes we are angry about recent things but this is no way to act

... feminazis don’t represent me, it’s about fighting for the end of violence against women, not about becoming wild beasts #TheyDon’tRepresentMe

You think you’re so smart, but this tweet proves you’re not. Don’t confuse things. Nothing justifies these acts of vandalism that do nothing to help. On the contrary, they only discredit the movement. There are smarter ways of demonstrating #TheyDoNotRepresentMe



**Figure 3.** Dendrogram of #TheyDoNotRepresentMe own elaboration.

By calling the women who protested “animals” or “wild beasts” these tweets refer to the affective binarism that renders visible that politics is understood as a rational exercise and therefore contrary to affectivity (Bargetz, 2015). In addition to the first two tweets, several other tweets call upon women to demonstrate in a more “intelligent” way, which in the context of the affective polarization that emerged from the data refers to not showing affects, such as rage, on historical monuments. The last tweet in particular shows that intelligence is related to non-affective forms of protest. Hence, the women who participated in the protest are portrayed as unintelligent, given that they could not control their affects. This marks a clear division between the “intelligent” way of demonstrating, understood as the “rational” one, and the “animal” one associated with women. Likewise, these tweets reinforce the idea that women are more affective and therefore can be compared to animals or wild beasts, as a way of implying that they do not have political reasoning. These findings are supported by the feminist literature that gives an

account of affective binarism based on gender stereotypes (Liljeström, 2015).

In the second cluster, the conversation brought together diverse topics, such as “angel” (angel), “desmadre” (rampage), “exigir” (demand), “derechos” (rights), “sociedad” (society), “odio” (hate), “destrozos” (destruction), and “femicidios” (femicides). In this cluster, the relationship established between fear, hatred, and destruction stands out. It followed the same pattern as the first cluster. Several tweets evoked fear that generates violence against women but pointed out that this does not justify the violence of the protests that in turn generate more fear and can only result in hatred of feminism. Despite the fact that several tweets condemned the ways in which women showed their affections in the public arena, the same tweets mobilized affective patterns of how one should feel after a feminist protest. As in the first cluster, after evoking these affections, the threat as a form of disciplining becomes present. The following tweets show the tone of the discussion:

---

Do they really think there will be justice, equity, peace and respect acting with violence and vandalism? The only thing they incite is hatred, and there will be someone who will put a limit to them to beat\*, really take it down a notch #TheyDoNotRepresentMe

#TheyDoNotRepresentMe fucking women whores, put them to clean their mess, damned criminals

---

Finally, in the third cluster, there are words, such as “neta” (really), “locas” (crazy), and “pinches” (fucking). The word

“crazy” was present in several tweets as ways of referring to women protesters, as the following tweets show:



Destroying, vandalizing, assaulting, violating, mistreating. Definitely the Femi crazies at today's demonstration do not represent me. If they want respect, they should give respect. #TheyDoNotRepresentMe

fucking crazy women this would not be allowed by Putin imagine now a march of men we are going to kill them through fucking blows #TheyDoNotRepresentMe

They want respect and equality? Several of them deserve jail #TheyDoNotRepresentMe

These tweets account for corrective threats that incite violence from women protesters. The idea of “if they want respect, they must give respect” can be read as an affective negotiation that accounts for the way in which women are politically infantilized (Krook, 2020a). That is, respect for women must be earned through behaving “well” in public spaces, rather than seeing them as political subjects who activate their rights, such as protesting. In the tweets, there is evidence that this respect and equality is subject to an exchange: If they “behave well,” we respect their right to protest. On the contrary, if they “misbehave,” women are

deserving of beatings or even jail as mechanisms of political disciplining.

In Figure 4, the existence of four clusters can be observed. In the first one, the users had a conversation about the monument Angel de la Independencia (historical monument), “revolution” (revolution), and “ninguna” (no one). In this cluster, the conversation was around the different ways in which women impacted history. It was said that no revolution had ever been peaceful. As a way to incite a feminist revolution, the claim was not only to exercise the right to protest but also the potential of their affective politics. Please see the following tweets:

I don't remember independence or revolution being peaceful, but because they were led by men they were called heroes. Really still don't see it? #TheyDoRepresentMe

#TheyDoRepresentMe When tyranny is law, revolution is order. 🇺🇸❤️

THE REVOLUTION WILL BE FEMINIST. #TheyDoRepresentMe and represent all those who no longer have a voice. If you are not going to support, don't get in the way!

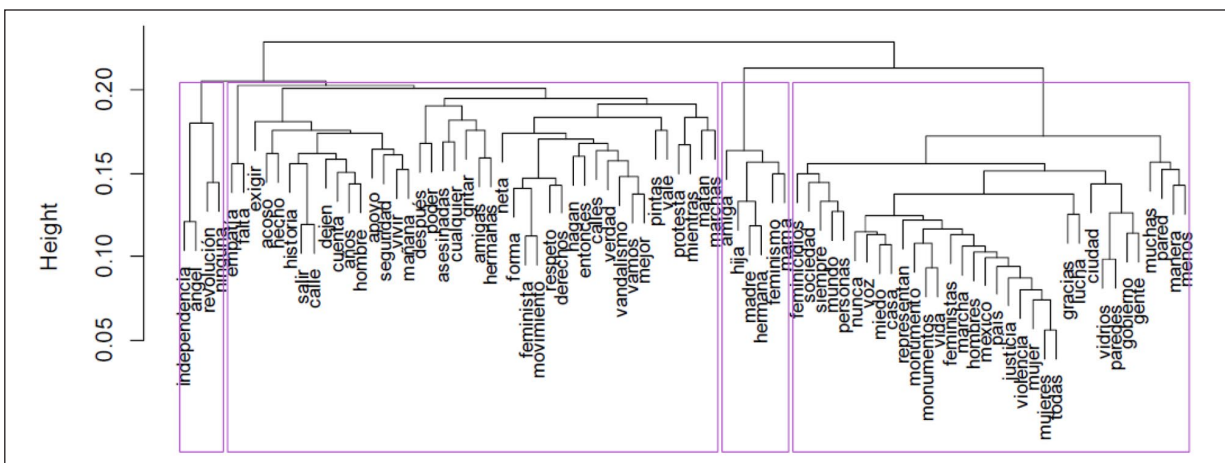


Figure 4. Dendrogram of #TheyDoRepresentMe.

The first tweet refers to the different ways in which gender and emotions are socially distributed (Bargetz, 2015). While the strength, the rage to raise one’s voice and challenge the public order is seen as part of a masculine agency that has historically made men out to be as heroes, women are disciplined and called denigrating words for disrupting the public order. This is why women subvert the word “revolution” by proposing feminist version of it.

In the second cluster, discussions prevailed about the severe situation of violence toward women. The key topics were “gritos” (shouts), “acoso” (harassment), “seguridad” (security), and “respect” (respect). This cluster was close to the third in meaning where the words women, daughters, mother, female friend appear as ways of naming the different stages of life for women who are affected daily by violence. As a feminist strategy to return the focus of the

discussion to the serious situation of violence against women, the affectivity of violence was discussed in both clusters as the need to scream to be heard. A pattern that was clear in the conversation was the blame they placed on the

government for impunity. When referring to discussions about representation, they reaffirmed that the women protesters did represent them and not the government, which does not care about violence against women.

---

I don't feel "represented" with a government that lets women be raped and killed every day and does nothing #TheyDoRepresentMe

When no one listens to you have to shout. #TheyDoRepresentMe

The dudes who rape, kill, abuse and discriminate against women are the same dudes who demand that you behave well and not make a mess, because "those are not manners". Go fuck yourselves. #TheyDoRepresentMe

---

In the last cluster, it was discussed how graffitied windows and walls generated more outrage than feminicides. Several tweets discussed the fear that women suffer every day of not returning home. They put the focus on the situation of violence that women in Mexico have to live with on a daily basis. They also talked about the government protecting the walls more than the women. These particular tweets made reference to the police security operation mobilized by the government before 8 March 2020 to protect historical monuments. They also pointed out that society gives more importance to monuments and walls than to the 10 women who are

murdered every day in Mexico. The distribution of gendered affect was also brought up in the sense that women were lynched in social networks for painting walls or historical monuments, meanwhile, the affectivity of male soccer-goers that also intervene at historical monuments and vandalize public spaces is met with comparative silence. Therefore, women mobilized the phrase "I prefer to see a graffiti than a dead woman" or hashtags, such as #MujeresNoParedes (WomenNotWalls) as forms of counterargument to the hate speech due to graffiti on historical monuments. The following tweets give evidence of this discussion:

---

Oh, how I wish to be wall for you to be outraged if they touch me without permission! 🙄 #TheyDoRepresentMe

Men going to insult feminists for painting a wall or a monument asking for justice after they break everything for a soccer match. #TheyDoRepresentMe

I don't give a shit about monuments, why do I want to live in a fucking country with monuments and beautiful walls if it's a feminicidal country 🙄 #TheyDoRepresentMe

---

These tweets testify to the how objects are catalyzers of affects (Bennett, 2010) and the polarization that this provoke. On one hand, #TheyDoNotRepresentMe shows that the historical monuments provoked an affectivity against the protesters even stronger than the violated bodies of women. On the other hand, the #TheyDoRepresentMe shows that the murdered bodies of women have evoked solidarity a new feminist agency that has brought women all over Latin America to the streets (de Souza, 2019; Hemmings, 2012). However, despite the fact that #TheyDoRepresentMe to vindicate the feminist protest through putting the focus on the protest, comments against the protest were hung on that hashtag, too, to continue denigrating and violating women. In fact, there are several tweets that, although they are in #TheyDoRepresentMe, have the same tone of condemnation and disciplining women for their ways of protesting.

## Discussion and Conclusion

The results of this study showed that an affective polarization followed the feminist protests in Mexico in 2019 and

2020 that was visible in the analyzed hashtags. On one hand, the hashtag #TheyDoNotRepresentMe rendered visible the affective disciplining (Savigny, 2020) directed to feminist protesters. The tweets tend to instrumentalize affects, such as shame and hate, against the feminist mobilization to create an enemy—"they," the ones that are violent, versus "us," the defenders of our city. It was also observed that the tweets also established shared perceptions (Ahmed, 2014) of how "we" should feel about a feminist protest: fear, shame, and disgust.

The tweets mobilized affective mechanisms as they reinforce gendered stereotypes, such as women being crazy, emotional, furious, or animals (Bargetz, 2015), as a strategy to polarize public opinion. Politically active women were recast in the aftermath and constructed as violent, out of control, and in need of discipline. These outcomes agree with existing literature on the ways in which women's affective agency is denied (Alison, 2004). However, the tweets were not limited to discriminatory adjectives and hate speech against feminism. As the data revealed, more than the half the dataset showed a negative and very negative sentiment

and some of them high levels of toxicity. In the article, this toxicity presented evidence of incitation of (sexual) violence and murder. This also shows how women were infantilized and even their political right of protest was subject to negotiation of respect dependent on their behavior. On the other hand, the hashtag #TheyDoRepresentMe had the objective of changing the focus from the supposed “violence” of the protests toward the one that men exercised against women. With this hashtag, women rendered visible the diverse affectations of violence and impunity in their lives, and they also politicized their affects rage, hope, and solidarity as a renewed way of doing feminist politics (Hemmings, 2012).

Our cluster analysis confirmed the affective polarization that surrounds the violence of the protest versus the violence against women. It points out the ways in which women were disciplined for protesting whereas men may disrupt public order and are not condemned for it—or if they are, not with threats of sexual violence. This finding is supported by theoretical discussions on heteronormative conceptions of femininity that reinforced the idea that women are not supposed to be violent (Åhäll, 2016, 2018; Alison, 2004).

Another point of polarization was the fear and hate against feminism versus the rage against government and men as the actors and accomplices of gendered violence. There was also a polarization about the politics of representation of women protesters versus the representation of and uncaring government that does nothing against gendered violence. Added into the mix was the condemnation of protest against women’s right to protest. All these issues that were subject to polarization showed the affective mechanisms to control the way women get involved in political life. They also showed the reinforcement of reproduction of affective stereotypes of women as highly affective, without rationality. Moreover, the polarization reinforces the existence of the binarism between affects and politics (Liljeström, 2015).

Although the condemnation of protests is not reserved only for women, here, we can observe the affective gendered aspect of violence to silence women in the political arena. For instance, women protesters were attacked in relation to traditional gender roles through tweets claiming that women should remain in the kitchen instead of protesting or that they need to clean all the mess that they left behind on the streets. While various other tweets incited to sexual violence. This result shows the mobilization of constructions of gender in social networks reproduce classic places for women out of politics. We claim that the disciplining and punishment toward feminist protesters is just another attempt to control and subordinate women’s bodies to victims without agency which at the same time makes visible how patriarchy is manifested on social media.

All in all, we can confirm our argument that the affective polarization aims at disciplining the affective political participation of women. With this, while it is true that social media has had an impact in amplified feminist mobilization, it has had also counter-implications on the increase of digital

violence through affective governance mechanisms after feminist mobilizations. We believe that this toxic digital violence has implications for women’s political participation writ large: for example, the creation of a stereotype of feminist protesters as crazy and furious women who have to be disciplined through sexual or police violence. These could intensify the crisis of human rights in Mexico through stigmatization, violence, and criminalization of feminist protests (Amnistía Internacional, 2021). Due to the fact that the most toxic mechanisms are the incitation toward murder and sexual violence, there is a need of more studies that analyze both digital and non-digital violence, and their potential correlation since toxicity on social media has gone beyond internet. Based on our study, we believe that social media is a networked affective space (Hiilis et al., 2019) that break downs barriers for continuing oppressing women.

Another important finding that should be further explored is the materiality of affectivity (Ahmed, 2014; Bennett, 2010). At the core of this polarization is the affectivity that emerged from the graffitied monuments versus the female bodies as a subject of dispute. How historical monuments evoke strong affects against feminist protesters, and how feminists simultaneously politicized the affectivity toward violence against female bodies. This result suggests we must consider the symbolic-discursive affective polarization seen on Twitter in a broader way to render visible the linked materiality between objects and bodies in the digital society. This could be considered as a study that points out possible further research on how affects are instrumentalized to sustain patriarchy on social media and at the same time how the affective feminist politics are mobilized to dispute this.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Marcela Suarez acknowledges the support of the research group DiGiTAL funding by the Berlin Equal Opportunities Program.

### ORCID iD

Marcela Suarez Estrada  <https://orcid.org/0000-0003-2412-5214>

### Note

1. Sor Juana de la Cruz was an historical female nun. She was a writer, religious, educated, and a cultural canon in Mexico.

### References

- Åhäll, L. (2016). On the imagination of “Woman” as killer in war. In S. Sharoni, J. Welland, L. Steiner, & J. Pedersen (Eds.), *Handbook on gender and war* (pp. 47–65). Edwar Elgar.

- Åhäll, L. (2018). Affect as methodology: Feminism and the politics of emotion. *International Political Sociology*, 12(1), 36–52.
- Ahmed, S. (2014). Afterword: Emotions and their objects. In S. Ahmed (Ed.), *The cultural politics of emotion* (pp. 204–233). Edinburgh University Press.
- Alison, M. (2004). Women as agents of political violence: Gendering security. *Security Dialogue*, 35(4), 447–463. <https://doi.org/10.1177/0967010604049522>
- Amnistía Internacional [Amnesty International]. (2021). *México: La era de las mujeres: Estigma y violencia contra mujeres que protestan*. <https://amnistia.org.mx/contenido/wp-content/uploads/2021/03/VF-Mexico-La-Era-de-las-Mujeres-FINAL.pdf>
- Ananías Soto, C., & Sánchez Vergara, K. (2019). Violencia en Internet contra feministas y otras activistas chilenas. *Revista Estudios Feministas*, 27(3), 1–13.
- Arjona Estévez, J. C. (2019). *Informe sobre las violencias de género en la procuración de justicia en la Ciudad de México*. [https://cdhcm.org.mx/wp-content/uploads/2019/09/Informe\\_violencia\\_de\\_genero.pdf](https://cdhcm.org.mx/wp-content/uploads/2019/09/Informe_violencia_de_genero.pdf)
- Bargetz, B. (2015). The distribution of emotions: Affective politics of emancipation. *Hypatia*, 30(3), 580–596.
- Bennett, J. (2010). *Vibrant matter. A political ecology of things*. Duke University Press.
- Davis, S., & Santillana, M. (2019). From the streets to the screen to nowhere: Las morras and the fragility of networked digital activism. *Westminster Papers in Communication and Culture*, 14(1), 18–32.
- Deleuze, G., & Guattari, F. (1987). *A thousand plateaus capitalism and schizophrenia*. University of Minnesota Press.
- de Souza, N. M. F. (2019). When the body speaks (to) the political: Feminist activism in Latin America and the quest for alternative democratic futures. *Contexto Internacional*, 41(1), 89–112.
- Döveling, K., Harju, A. A., & Sommer, D. (2018). From mediated emotion to digital affect cultures: New technologies and global flows of emotion. *Social Media + Society*, 4(1), 1–12.
- Gallacher, J. D., Heerdink, M. W., & Hewstone, M. (2021). Online engagement between opposing political protest groups via social media is linked to physical violence of offline encounters. *Social Media + Society*, 7(1), 1–16.
- Hemmings, C. (2012). Affective solidarity: Feminist reflexivity and political transformation. *Feminist Theory*, 13(2), 147–161.
- Hiililä, K., Paasonen, S., & Petit, M. (2019). *Networked affect*. The MIT Press.
- Hynnä, K., Lehto, M., & Paasonen, S. (2019). Affective body politics of social media. *Social Media + Society*, 5(4), 1–5.
- Jupp, E., Pykett, J., & Smith, F. M. (2016). *Emotional states: Sites and spaces of affective governance*. Routledge.
- Kelly Garrett, R., Long, J. A., & Jeong, M. S. (2019). From partisan media to misperception: Affective polarization as mediator. *Journal of Communication*, 69(5), 490–512.
- Krook, M. L. (2020a). *Violence against women in politics*. Oxford University Press.
- Krook, M. L. (2020b). Violence against women in politics. In M. Sawyer, F. Jenkins, & K. Downing (Eds.), *How gender can transform the social sciences* (pp. 57–64). Palgrave Macmillan.
- Liljeström, M. (2015). Affect. In L. Disch & M. Hawkesworth (Eds.), *The Oxford handbook of feminist theory* (Vol. 1, pp. 16–38). Oxford University Press.
- Papacharissi, Z. (2016). Affective publics and structures of storytelling: Sentiment, events and mediality. *Information Communication and Society*, 19(3), 307–324.
- Pedwell, C., & Whitehead, A. (2012). Affecting feminism: Questions of feeling in feminist theory. *Feminist Theory*, 13(2), 115–129.
- Penz, O., & Sauer, B. (2020). *Governing affects*. Routledge.
- Reckwitz, A. (2016). How the senses organise the social. In M. Jonas & B. Littig (Eds.), *Praxeological political analysis* (pp. 68–78). Routledge.
- Ruiz, K. (2020, March 5). Autoridades investigan mensajes de odio. La Policía Cibernética rastrea post donde se llama a agredir a policieras en marcha del 8 de marzo. *El Universal*. <https://www.eluniversal.com.mx/metropoli/autoridades-investigacion-mensajes-de-odio>
- Sanín, J. R. (2020). Violence against women in politics: Latin America in an era of backlash. *Signs*, 45(2), 302–310.
- Savigny, H. (2020). *Cultural sexism: The politics of feminist rage in the #metoo era*. Bristol University Press.
- Sundén, J., & Paasonen, S. (2020). *Who's laughing now? Feminist tactics in social media*. The MIT Press.
- Ward, E., Crowhurst, I., & Sauer, B. (2019). Editorial affective governance and the sex trade. *Journal of Political Power*, 12(3), 313–317.
- Zarzycka, M., & Olivieri, D. (2017). Affective encounters: Tools of interruption for activist media practices. *Feminist Media Studies*, 17(4), 527–534.

### Author Biographies

Marcela Suarez Estrada is a Professor in Political Science at the Institute of Latin American Studies at Freie Universität Berlin. She holds a PhD in Political Science also from the Freie Universität. Her areas of specialization are sociopolitical dynamics of new technologies, governance, knowledge asymmetries, technofeminisms, and digital culture. Her current research project is titled “Feminist Politics and the Fight Against Violence in the Era of Digitalization” and is financed by the Berlin Equal Opportunities Program.

Yulissa Juarez (BSc Statistics) is a partial time collaborator at Centro de Estudios de Opinión y Análisis (CEOA) at Universidad Veracruzana, Mexico. Current research interests include data analysis and statistical computing.

C. A. Piña-García (PhD in Computer Science from University of Essex at the School of Computer Science and Electronic Engineering) is a full-time researcher in the Centro de Estudios de Opinión y Análisis (CEOA) at Universidad Veracruzana, Mexico. Current research interests include social network analysis, computational social science, and social mining.