Toxicity in Pro-CCP Tweets Targeting Prominent Regime Critics in the Uyghur Diaspora

Allison Koh[1], Jonathan Nagler[2], and Joshua A. Tucker[2]

[1]Hertie School

[2]New York University

December 13, 2022

Abstract

Chinese authorities and supporters of the Chinese Communist Party have leveraged Western social media platforms to lodge smear campaigns against Uyghur, Kazakh, and other Turkic women living abroad. The women targeted have spoken out against human rights violations targeting Muslim minorities in the Uyghur region of northwestern China. These state-sponsored social media attacks reflect a global trend in the use of digital tools for engaging in transnational repression. Against this background, this study investigates the gender dimensions and contours of pro-CCP smear campaigns on Twitter. We use large-scale quantitative data from Twitter and information on Chinese authorities' attacks targeting Uyghur individuals abroad to analyze whether women in the Uyghur diaspora are disproportionately attacked with vitriol on pro-CCP Twitter. Controlling for gender, we also empirically test whether content produced by overtly state-affiliated accounts differs from that of other pro-CCP users, and whether pro-CCP tweets in English differ from those published in Chinese. Finally, we explore the topics discussed in targeted mentions of Uyghurs in the diaspora and possible differences based on the gender of a targeted individual. Our findings have important implications for studying gender and state-sponsored disinformation from authoritarian contexts on Western media platforms.

Keywords: disinformation; digital transnational repression; authoritarianism; gender; Twitter

1 Introduction

Repressive governments are increasingly leveraging the affordances of social media to harass online opposition beyond their borders. Women and individuals with other marginalized identities are especially vulnerable to these attacks. While relevant studies have provided much insight on how states leverage social media to influence politics at home and abroad, and how targeted harassment has been deployed in a handful of cases, little is known about the overall contours and possible gendered aspects of state-aligned content that targets the online opposition—especially in the transnational dimension. This study addresses this gap by identifying of how pro-CCP Twitter accounts communicate with regime critics in the Uyghur diaspora, whether women are targeted with more negative rhetoric, the extent to which official state accounts participate in targeted smear campaigns, and whether language could be strategically manipulated in media attacks.

In Section 2, we outline hypotheses and research questions. In Section 3, we present the research design of this project. We elaborate on variable measurement in Section 4. Our plan for analysis is detailed in Section 5, and we discuss ethical considerations for this study in Section **??**.

2 Theoretical Framework

The overarching question guiding this study is, *Are Uyghur women diaspora activists more likely to be targeted in pro-CCP smear campaigns on Twitter?* Our primary hypothesis tests the extent to which women activists in the Uyghur diaspora targeted with toxic language on pro-CCP Twitter:

> **Hypothesis 1** *In pro-CCP content on Twitter, we expect that women diaspora activists are more likely to be targeted with toxicity than men diaspora activists.*

We also test the differences between tweets authored by officially state-affiliated accounts and those from other pro-CCP users. In the offline world, violent repression is often outsourced to "thugs-for-hire" to provide the state with some plausible deniability regarding their involvement in specific attacks (Ong 2020). In the digital age, repressive states outsource the production of pro-regime content to curate the illusion of having their positions "endorsed" by seemingly independent voices (Nyst and Monaco 2018). In line with the provision of these repressive tactics that take place online and offline, we generally expect that official state-affiliated accounts use less violent rhetoric

when targeting the online opposition. Accordingly, we expect the following differences in content produced by state-affiliated accounts:

> **Hypothesis 2** *Controlling for gender, tweets by state-affiliated accounts are less likely to be toxic towards activists compared to other pro-CCP users.*

Our final hypothesis tests whether we observe variation in the toxicity of posts by the language that they are posted in. In general, English-language tweets represent content that is more easily accessible to international audiences. Tweets in other languages are generally less accessible to a global audience, and might broadcast different messages accordingly. In the Uyghur context, Mandarin Chinese represents the language of the oppressor. The predominant language in the Uyghur region is Uyghur, which is a Turkic language similar to Uzbek but written in Arabic script. However, since the early 2000s, Mandarin Chinese has been imposed on the Uyghur population as part of their strategy for escalating repression in the region. With this variation in purported audiences and the role of Mandarin Chinese as a repressive language in the context of this study, we expect to observe the following:

> **Hypothesis 3** *Controlling for gender, English-language tweets on pro-CCP Twitter are less likely to be toxic towards activists compared to tweets published in Chinese.*

We also aim to identify the diversity of topics discussed in pro-CCP tweets that target Uyghur diaspora activists. To better understand how pro-CCP Twitter attacks Uyghur diaspora activists, and possible variation by the gender of a targeted activist, we address the following research questions:

> **Research Question 1** *Which topics appear in pro-CCP smear campaigns targeting Uyghur diaspora activists on Twitter?*

> **Research Question 2** *Do pro-CCP accounts discuss different topics when targeting women Uyghur diaspora activists, compared to when they target men?*

3 Data Collection and Preparation

3.1 Data Sources

In this study, we draw from multiple sources of data, summarized in Table 1. Information on prominent regime critics in the Uyghur diaspora are obtained from the China's Repression of the

Uyghurs (CTRU) dataset. We use the full-archive search endpoint from the Twitter Academic API to collect tweets mentioning (by name with/without spaces and username) prominent regime critics. We collect information on Chinese state-affiliated Twitter accounts from the Alliance for Securing Democracy's Hamilton 2.0 Dashboard. We plan to append this list of 250 accounts associated with Chinese government officials, state-backed media outlets, and diplomatic entities to gain a more comprehensive overview on the activity of overtly state-affiliated actors in the context of this study.

Table 1: Summary of Data Sources

Description	Time Frame	Source
CTRU Dataset	2002-2021	Lemon, Jardine and Hall (2022)
Twitter mentions of regime critics	2017-2022	Twitter Academic API
Chinese state-affiliated Twitter accounts	N/A	Alliance for Securing Democracy (2022)

We evaluate the extent to which women Uyghur activists residing outside of China are targeted on pro-CCP Twitter by analyzing the tweets that mention Uyghur activists abroad. Information on targeted individuals in the Uyghur diaspora will be drawn from a gender-balanced sample of individuals included in the data on incidents of China's transnational repression targeting the Uyghur diaspora[1]. Gender is inferred from the pronouns used to describe the targeted individual in the description of an incident in the database on transnational repression, or in cited sources (=1 if she/her pronouns are used, 0 otherwise). Our sampling strategy for identifying individuals in the CTRU data is elaborated in Section 3.2. We also use these data to collect auxiliary information on targeted individuals' host countries and political activity.

Our primary source of social media data is the Twitter Academic API's search endpoint for collecting historical data from all publicly available tweets. In particular, we queried all tweets mentioning the names and Twitter handles in our input list between January 1, 2017 and April 30, 2022. We use January 2017 as the starting point of our data collection to encompass tweets that were published around the time that the Chinese government enacted "XUAR De-Extremification Regulation" legislation in early 2017. This legislation had important implications for how foreign governments treated Uyghur and Turkic Muslims trying to flee the region. For instance, potential

[1]Information on these incidents were collected from searches in topically relevant reports from human rights organizations (e.g. Amnesty International, Human Rights Watch, World Uyghur Congress, and the Uyghur Human Rights Project), keyword searches from Radio Free Asia and newswires, and existing datasets on global transnational repression (Jardine, Lemon and Hall 2021).

escape routes for Muslim minorities throughout Asia and the MENA region were dwindling (Jardine 2022). Other notable events that occurred within the data collection period include the Trump administration's declaration of genocide in the region (Pompeo 2021) and the 2022 Beijing Olympics, also referred to as the "Genocide Games" by individuals around the world who boycotted the Olympics because of human rights abuses committed by the Chinese government in the region. The resulting dataset comprises of all potentially relevant tweets, from which we identify pro-CCP content.

3.2 Sampling Strategy

The population of interest in this study comprises of any regime critic in the Uyghur diaspora who advocates for the rights of Muslim minorities in the Uyghur region and is active on Twitter. In this study, we primarily focus on Uyghur activists who are in exile, but plan to expand our analysis to more broadly understand pro-CCP tweets engaging with any Uyghur advocate who shares a common heritage with Muslim minorities in the Uyghur region. In the context of this study, we establish that an individual is part of the Uyghur diaspora if (i) they are included as a target of incidents recorded in the CTRU data or (ii) they indicate it in their username or profile description[2].

For our primary analyses, we study pro-CCP tweets, published in English and Mandarin Chinese, which mention a gender-balanced sample of 22 individuals who are active on Twitter and have been targeted in more than one publicly recorded incident of Stage 1 Chinese transnational repression, which comprises of non-physical forms of coercion[3]. This sampling strategy allows us to focus on attacks targeting more "prominent" activists who are targeted with tactics who have experienced repression, but not in incidents so severe that they are being silenced through physical coercion. The findings from tweets mentioning this smaller sample of individuals are important for exploring the broader implications of our findings. To ensure that results are consistent beyond tweets mentioning the individuals in this sample, we plan to run robustness checks on larger gender-balanced samples drawn from the CTRU dataset.

[2]We are discussing the implications of expanding our analyses beyond the initial sample of 22 prominent Uyghur advocates in the diaspora, with a focus on the trade-off between the ethical principles of beneficence and respect to persons. While we may have to impose additional scope conditions to our research if we solely rely on expert-compiled datasets, using alternative methods of identifying "potential" targets of Chinese repression via snowball sampling on Twitter could put users who are not already in the public eye at additional risk of harm.

[3]Relevant incidents include warnings and threats to individuals and/or their family members, as well as arrest requests issued bilaterally or multilaterally (e.g. via INTERPOL).

3.3 Data Annotation Process

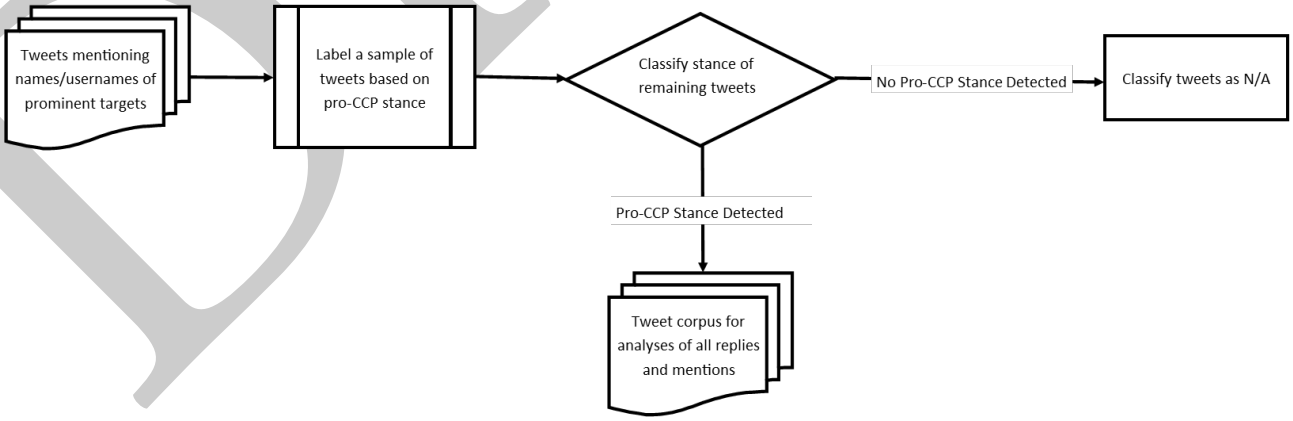
Our data annotation pipeline for labeling tweets in English and Chinese is illustrated in Figure 1. With names[4] and Twitter handles as inputs for collecting social media data, we filter out irrelevant content by training crowd-sourced annotators to classify a gender- and language-balanced sample of 5000 tweets mentioning Uyghur diaspora activists. For each tweet, three annotators label tweets based on whether they represent a pro-CCP stance. We indicate that typical clues or signals include:

- Cites justifications for CCP actions and policies
- Questions the knowledge or experience of individuals negatively commenting on CCP policies
- Includes derogatory or insulting language towards Chinese ethnic minority groups
- Promotes calls of unity of different groups

We also note in our instructions that references to the Uyghur region as "Xinjiang" (the CCP's official name for the region) are not always indicative of a pro-CCP stance. Further, we clarify that tweets referencing pro-CCP content may indicate a neutral stance, which is irrelevant to this study.

Tweet-level information on posts labeled as pro-CCP is shown in Table 2. Of the 5000 tweets in this sample, the coders indicated that 534 (10.68%) of tweets represented a pro-CCP stance. Using Fleiss's kappa measure of intercoder agreement, the English sample of labeled tweets demonstrates a higher level of agreement (α = 0.536) compared to the Chinese sample (α = 0.213). In cases of disagreement, we designate a tweet as pro-CCP if 2 coders (67%) identify it as such.

Figure 1: Data generation and annotation workflow for primary analyses



[4]Full names are queried with and without spaces, to include tweets that mention individuals without tagging them (e.g. in a hashtag).

Table 2: Information on tweets labeled as positive

		N	%
Agreement %	67%	339	63.5
	100%	195	36.5
Language	English	319	59.7
	Chinese	215	40.3

Once we subset for relevant content, we augment our data with additional information on toxicity and the users producing pro-CCP content. We use the Perspective API to derive information on the probability that each tweet is *toxic*, or the probability that a tweet will lead to a target limiting their activity on the platform. Variable measurement with these tweet labels are elaborated in Section 4.1. We supplement this final set of tweets with information on which accounts are state-affiliated, based on a list of government/diplomatic and state-backed media accounts tracked by the Alliance for Securing Democracy (2022) Hamilton 2.0 Dashboard[5]. We elaborate on how these labels will be used for measurement and analysis in Sections 4 and 5.

4 Variable Measurement

4.1 Dependent Variable

Our outcome measure of interest is the probability that a tweet is toxic. A tweet is *toxic* if it has the propensity for a target to limit their social media activity, which in the this study pertains to how Uyghur diaspora activists share their perspectives on human rights abuses in the Uyghur region. The probability that a tweet is toxic is measured by the Perspective API.

4.2 Independent Variables

Gender is the primary independent variable of focus for testing all hypotheses. It is measured as a binary variable, coded as 1 if she/her pronouns are used to refer to a Uyghur diaspora activist in detailed notes or cited sources from the dataset compiled by the Oxus Society for Central Asian Affairs. We presume that the reference group for this variable is predominantly male. The other independent variable of interest is whether a tweet author is a *state-affiliated account*. A tweet is

[5]We are working on appending this list of state-affiliated accounts to obtain a comprehensive view of which accounts overtly affiliated with the Chinese government are targeting individuals in the Uyghur diaspora.

authored by a state-affiliated account if the user appears on our list of overtly affiliated government, diplomatic, or state-backed media accounts. Our final independent variable of interest is *language*, which include English and Chinese.

5 Empirical Strategy

5.1 Hypothesis Testing

All analyses are performed at the level of individual *mentions*, as a single tweet could include mentions of multiple individuals in our sample of Uyghur regime critics residing abroad. We plan to use multiple linear regression analysis to assess the difference in average toxicity scores based on our independent variables of interest. As specified in Section 4.1, the outcome variable of interest for hypothesis testing is the probability that a tweet is *toxic*. The primary independent variable of interest for all hypotheses is the *gender* of a targeted Uyghur diaspora activist, with the aim of comparing differences in average toxicity scores between women and men. We initially run all specifications without control variables, but also run specifications with controls. Controls on the individual level are derived from newswires and datasets on global transnational repression, while country-level variables are derived from human rights reports on the repression of Muslim minorities in the Uyghur region.

To test our first hypothesis, our primary independent variable of interest is gender. For our second hypothesis, our independent variables of interest are gender and whether a tweet is authored by a state-affiliated account. For our third hypothesis, our independent variables of interest are gender and the language that a tweet is published in. We conduct primary analyses with a corpus of tweets that mention 22 prominent Uyghur diaspora activists who are active on Twitter.

5.2 Exploratory Analysis

In addition to the aforementioned hypothesis testing, we engage in exploratory analysis to investigate possible sources of variation in the topics observed in pro-CCP smear campaigns that target Uyghur diaspora activists. We also investigate whether we observe differences in the topics used in tweets that target women diaspora activists compared to men diaspora activists.

5.3 Addressing Limitations

Once we complete our primary analyses, we aim to address limitations in our data generating process, variable measurement, and the scope conditions of this research.

We aim to correct for data generating bias by seeking out sources of Twitter data that are not procured using the full-archive search endpoint of the Twitter Academic API. A potential limitation with solely using historical data from the Twitter Academic API's full-archive search is that we are missing tweets that have been removed from the platform. As a robustness check, we initially planned to compare the results of our analyses with deleted tweets from Twitter's Information Operations archive. However, upon inspection of the relevant releases from these data, we only found eight tweets relevant to this study and have decided not to include them in analysis. Our options outside of historical data are thus limited, but the implications of solely using historical data of tweets that remain on the platform remain important. While we do not have access to pro-CCP content that has been deleted, identifying what *does* remain on the platform is crucial for understanding how content moderation of social media posts that align with repressive governments can be improved in the future.

We also plan to adjust for variable measurement bias, particularly in how the Perspective API toxicity scores are measured. Recent research on Perspective and other toxicity classifiers has highlighted how toxicity scores are higher when words tied to marginalized identities (e.g. woman, gay) are present (Reichert, Qiu and Bayrooti 2020). We also speculate that the Perspective API performs differently across languages. To address this potential issue in gender bias, we plan on taking a sample of tweets, replacing words referring to women with words referring to men (e.g. swapping "she/her" with "he/him")—and vice versa—to compare toxicity scores of the original tweets to the altered set of tweets. We also plan on comparing the differences in toxicity scores of tweets across languages with a machine-translated altered sample of tweets. We plan on including any observed bias in gendered words and language as error terms in our analyses. To address other concerns with the Perspective API, we also plan on running our models with the dependent variable measured with other toxicity classifiers.

Finally, we address issues of selection bias and limited scope in our sample of individuals for primary analyses. Inferences drawn from focusing on 22 prominent Uyghur diaspora activists who are active on Twitter may limit the broader applicability of this research to all Uyghur diaspora

activists who are targeted in pro-CCP smear campaigns on Twitter. We are currently working on a strategy to expand our data collection without exposing individuals in the Uyghur diaspora who are not as visible in the public eye to potential harm.

6 Preliminary Analysis

6.1 Identifying pro-CCP tweets in English

We use the labeled sample of 5000 tweets to train text classifiers for detecting pro-CCP stance. As an initial foray into this challenge, we trained BERT and DistilBERT models as a baseline, with a train/test/validation split of 70/15/15. Thus far, we trialed training BERT and DistilBERT models sequentially with different learning rates (LR = 0.00001, 0.0001, 0.001). Of all six model specifications, two models yielded results that included labels in the pro-CCP category. The results for these two models (BERT with LR = 0.00001 and DistilBERT with LR = 0.0001) are shown in Table 3. With the F1-score as our basis of assessing model performance, we find that the DistilBERT model is our best-performing model and baseline to iterate on (F1 = 0.41). Figures 2 and 3 demonstrate the performance of the DistilBERT model and point to areas of improvement for architecture fine-tuning.

Our next steps for improving the performance of English-language classifiers are to test different modifications for fine-tuning the architecture of our baseline BERT models and addressing issues of class imbalance. Regarding the latter, we plan on combining our best-performing BERT model with Active Learning to improve the performance of our classifier.

Table 3: Baseline results from English-language text classifiers

Model	Precision	Recall	F1-Score
BERT	0.37	**0.42**	0.39
DistilBERT	**0.50**	0.35	**0.41**
RoBERTa	0.37	0.42	0.39

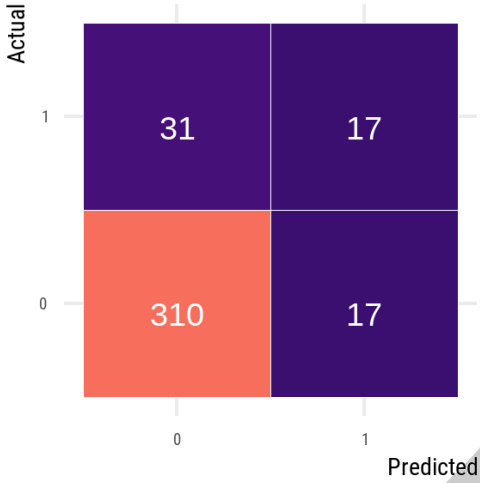


Figure 2: Confusion matrix of results from the DistilBERT classifier

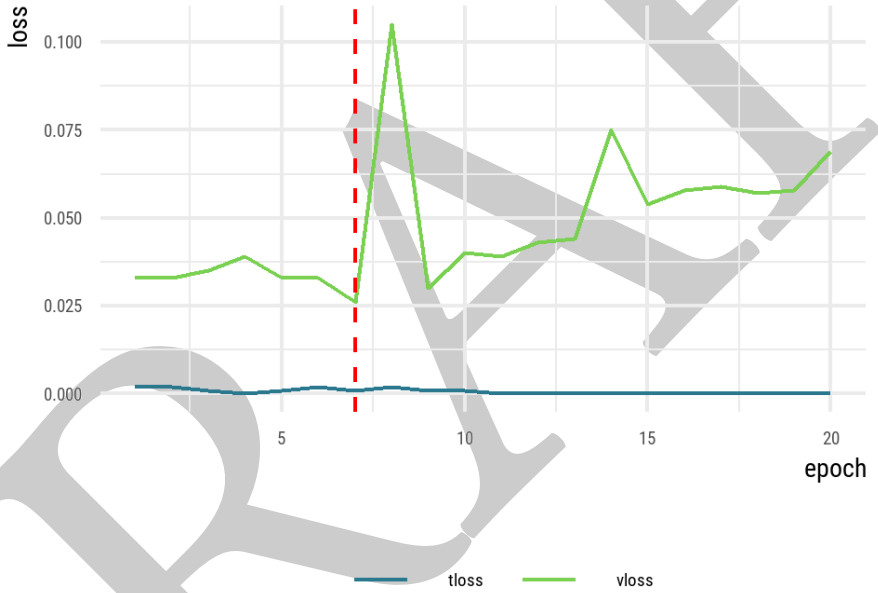


Figure 3: Training and validation loss from the underfit DistilBERT model. The red line represents the model used to yield predictions on the test set for this analysis.

6.2 Next steps

Once we have optimized the performance of our classifier on English-language tweets, we plan on tackling the challenge of classifying the Chinese-language tweets by stance. Once the tweets are labeled, we will transform the dataset to represent individual mentions as the unit of analysis for hypothesis testing. We will then carry out the robustness checks outlined in Section 5.3.

7 Concluding Remarks

In this study, we aim to understand the gender dimensions and contours of pro-CCP content on Twitter, with a focus on tweets mentioning prominent regime critics in the Uyghur diaspora. The findings from this research has important substantive implications for our understanding of the "grey areas" of state-aligned content that does not get removed from platforms. From a methodological standpoint, we introduce an empirical basis of studying targeted, state-aligned attacks in a multilingual context on Twitter.

References

Alliance for Securing Democracy. 2022. "Hamilton 2.0 Dashboard." https://securingdemocracy.gmfus.org/hamilton-dashboard/.

Jardine, Bradley. 2022. "Great Wall of Steel: China's Global Campaign to Suppress the Uyghurs." *Wilson Center* .

Jardine, Bradley, Edward Lemon and Natalie Hall. 2021. "No Space Left to Run: China's Transnational Repression of Uyghurs." *Uyghur Human Rights Project and Oxus Society for Central Asian Affairs* .

Lemon, Edward, Bradley Jardine and Natalie Hall. 2022. "Globalizing minority persecution: China's transnational repression of the Uyghurs." *Globalizations* pp. 1–17.

Nyst, Carly and Nick Monaco. 2018. State-Sponsored Trolling: How Governments Are Deploying Disinformation as Part of Broader Digital Harassment Campaigns. Technical report Institute for the Future.

Ong, Lynette H. 2020. *Outsourcing Repression: Everyday State Power in Contemporary China.* Oxford University Press.

Pompeo, Michael R. 2021. "Press Release: Determination of the Secretary of State on Atrocities in Xinjiang." \url{https://2017-2021.state.gov/determination-of-the-secretary-of-state-on-atrocities-in-xinjiang/}.

Reichert, Elizabeth, Helen Qiu and Jasmine Bayrooti. 2020. "Reading between the demographic lines: Resolving sources of bias in toxicity classifiers." *arXiv preprint arXiv:2006.16402* .