HARASSMENT IN TIKTOK COMMENTS

A Pilot Test of the TikTok Research API

CYBERBULLYING RESEARCH CENTER

Sameer Hinduja, Ph.D.

Justin W. Patchin, Ph.D.

cyberbullying.org

June 30, 2023

Table of Contents

Introduction

TikTok is a leading short-form video platform with approximately 150 million monthly active users in the United States.[1] Users upload customized video content covering a wide range of topics, including entertainment, sports, hobbies, fashion, makeup, politics, and social issues, and often employ filters, hashtags, popular songs, and amusing quotes. They interact with each other through common social media features including likes, comments, and direct messages. The platform has gained widespread popularity due to its personalized video feed which quickly and impressively learns the types of videos that each user prefers, various featured viral trends and challenges that have arisen over the years, and its rising position in the social zeitgeist.

In the Spring of 2023, TikTok released a Research API to academic researchers in the United States in an effort to "enhance transparency with the research community" and "stay accountable to how we moderate and recommend content."[2] According to their documentation, approved access allows for the retrieval of:

- Public account data, such as user profiles, comments, and performance data, such as number of comments, likes, and favorites that the user receives
- Public content data, such as comments, captions, subtitles, and performance data, such as number of comments, shares, likes, and favorites that the video receives
- Public data for keywords search results

TikTok reached out to the Cyberbullying Research Center with a request to test-drive their new Research API. We were provided access on May 30th, 2023, and were to publish a report of our findings by June 30, 2023, prior to its deployment to other regions of the world. The goal was to put the Research API through a real-world research application to identify any potential barriers to researchers. We agreed to use the tool to examine a specific research question related to the extent of harassment that occurs in TikTok comments.

We need to be clear at this point that while TikTok asked us to put the new tool through its paces, they did not provide any guidance, feedback, support, input, or opinions as it relates to the process and outcomes of this project and this report. They were not made privy to any aspect or version of this report prior to its public release. We must also acknowledge that TikTok provided a monetary contribution to the work of the Cyberbullying Research Center in exchange for our efforts on this project, but again we completed this work independently of TikTok.

Research Question

To explore the Research API, we chose to examine the extent of harassment that occurs in comments to user videos. For the purposes of this analysis, we chose to focus on public figures. TikTok formally defines "public figures" as anyone 18 years of age or older with "a significant public role, such as a government official, politician, business leader, or celebrity."[3] When it comes to harassment, we were generally looking for any targeted expressions that would make a reasonable person feel uncomfortable, distressed, demeaned, or bothered. We were also interested in identifying instances of sexual harassment, hate speech, or threats of physical harm directed toward these public figures.

Public figures may be harassed more often than the average social media user because they have more exposure, influence, and followers. This may attract more attention, criticism, or envy from others when they express their perspectives.[4] They may also face more targeted or coordinated attacks from groups or individuals who have political, ideological, or personal motives to do harm towards them.[4]

Public figures may also have more resources, support, experience, and/or resilience to deal with such attacks. Many presumably have social media handlers, legal support staff, and public relations and communications teams ready and able to serve as a protective layer against online aggression from others. All of this said, the presence and activity of public figures on TikTok is arguably one of the main reasons why *private figures* (i.e., general members of society) use the platform: to connect with and hear from those individuals who have achieved success and who occupy elevated positions in their respective fields. Accordingly, it seems prudent for the platforms that invite and host their presence and participation to more fully understand what public figures face, and then support and enhance their experience so that they feel safe within the community.

For this pilot test, we identified 10 politicians and 10 celebrities that are among the most visible and active on TikTok. Some within the sample post more frequently than others, and one of the politicians has not posted new videos in 2023, but the goal was to focus on the public figures with a significant following. It should also be noted that all of the politicians sampled are Democrats. We were unable to find any Republican or third-party politicians with a meaningful TikTok presence.

Table 1. Public Figures Analyzed

	Username	Followers (as of June 15, 2023)	Total Comments to Ten Recent Videos[1]	Range per Video
Politicians				
Alexandria Ocasio-Cortez	@aocinthehouse	786.0K	57,298	422-24,000
Bernie Sanders	@bernie	1.4M	305	0-220
Gretchen Whitmer	@Biggretchwhitmer	210.3K	3,577	7-1,824
Cory Booker	@corybooker	380.6K	3,524	16-2,081
Gavin Newsom	@gavinnewson	328.3K	19,009	163-6,396
Ilhan Omar	@ilhanmn	248.4K	4,795	28-1,493
Jeff Jackson	@jeffjacksonnc	2.2M	172,483	3,020-37,400
John Fetterman	@Johnfetterman	239.3K	27,701	282-16,800
Jon Ossoff	@Jon	498.4K	37,852	247-9,155
Jamaal Bowman	@repbowman	216.4K	4,437	6-3,255
Celebrities				
Addison Rae	@addisonre	88.6M	24,615	343-5,844
Billie Eilish	@billieelish	47.9M	954,800	43,700-210,400
Charli D'Amelio	@charlidamelio	150.9M	46,058	1,416-10,300
James Charles	@jamescharles	37.9M	24,560	622-10,200
Kylie Jenner	@kyliejenner	53.2M	38,796	595-17,300
Lady Gaga	@ladygaga	9.0M	98,130	792-35,500
MrBeast	@mrbeast	83.8M	270,500	1,604-102-800
The Rock	@therock	71.1M	84,179	645-54,600
Will Smith	@willsmith	73.0M	23,977	117-8,625
Zach King	@zachking	77.4M	52,935	763-17,200

[1]According to review of videos on TikTok on June 15, 2023. Some of the total comments for each video is estimated/rounded due to how the number is displayed on TikTok (e.g., 14.2k).

Preparation of Data

The first step of this project involved using the Research API to acquire our data of interest. To connect to the Research API, we sent cURL requests via PHP, with the appropriate headers and request parameters as specified in the documentation. As we began our work, our intention was to use the API to fetch videos and comments associated with the public figure usernames listed in Table 1. However, we soon discovered that the API did not seem to return videos of exact matches of the username specified in the requests even when all the parameters sent were correct according to the API documentation (explained in detail below in the discussion on usernames in the API Limitations section of this report). This was a major obstacle in our process to acquire comments from the specific usernames we were targeting.

Ultimately, we abandoned this original plan to fetch videos by username (after trying multiple query variations).

Our new plan was then to manually select the ten most recent videos that each public figure had posted on TikTok. This involved visiting the profile of each public figure in our sample, selecting the most recent ten videos posted, and noting the video ID for each. After preparing a CSV file containing 20 TikTok video URLs (10 each for the ten politicians and ten celebrities), we edited our PHP script to work off of that file and to fetch comments from the Research API for each video ID specified. Since the Comments API had a limitation of a maximum of 100 comments per request, and a maximum of 1,000 comments per video ID, that meant that our script needed to iterate over comments in batches of 100 until the API would either stop giving further comments for that video, or until the 1,000 comment limit per video was reached.

The API runs did not go smoothly. There were repeated internal server errors returned by the API which interrupted and stalled the process. Furthermore, the daily limit of the API calls was much lower than anticipated and our quota was often filled before all the needed data were fetched.

Given a maximum limit of 1,000 comments per video, we expected a maximum of 200,000 comments for the 20 videos that we had selected. Most videos returned less than 1,000 comments. This resulted in 154,540 comments (57,706 across 10 videos from each of ten politicians, and 96,879 comments across 10 videos from each of ten celebrities) fetched during the window of June 15-17, 2023. This consequently served as our corpus of data for the pilot test.

Preprocessing steps included data validation, integrity checks, character encoding, tokenization, word boundary detection, and the removal of duplicate and blank comments. Given their nontrivial presence, this latter step merits further discussion here.

Duplicate Comments

There were numerous instances in the dataset where comments appeared to be duplicated. After further exploration we noticed that duplicate comments had the exact same id. We are not sure why these comments were duplicated in the dataset, but having the same id made them easier to reconcile. We counted 1,517 duplicates among the politician comments (2.6%) and 4,058 duplicates among the celebrity comments (4.2%). We are unsure why certain comments were duplicated, or why there were more (as a percentage) among the celebrity comments. Nevertheless, the duplicates were removed for subsequent analyses.

Blank Comments

It was also apparent that there were many blank comments. There were rows in the dataset that had id numbers, as well as date and time stamps, but no content within the text field. It seems reasonable to conclude that there had been a comment submitted or posted, but it was removed at some point (either by the creator or by TikTok). Knowing exactly what the blank comments represent is particularly important, especially for the purposes of this inquiry. It may be that certain comments were removed by TikTok due to a policy violation or by the creator for some other reason.

Curiously, there was significant variation among the creators reviewed when it came to blank comments (see Table 2). While @jeffjacksonnc, @Jon, and @charlidamelio had fewer than one tenth of one percent of their comments blank, @bernie had nearly three-quarters of his comments blank. This wide variation suggests some nonrandom cause of the blank comments. Perhaps @bernie is removing many more comments than the others, or those who are commenting on his videos are more likely to violate community standards and have their comments automatically filtered out. It is important that TikTok provide details on what the blank comments may signify (more on this below).

After removing duplicate and blank comments, we were left with a dataset which included 141,892 comments (92,042 celebrities and 49,850 politicians; see Table 3).

Table 2. Blank Comments in Dataset

	Blank Comments	Percent of Comments in Dataset
Politicians		
@aocinthehouse	1,404	14.0%
@bernie	987	74.1%
@Biggretchwhitmer	669	18.8%
@corybooker	545	17.3%
@gavinnewson	1,020	14.3%
@ilhanmn	261	6.4%
@jeffjacksonnc	4	<0.1%
@Johnfetterman	996	13.7%
@Jon	5	<0.1%
@repbowman	447	16.9%
Celebrities		
@addisonre	663	7.1%
@billieelish	275	2.8%
@charlidamelio	1	<0.1%
@jamescharles	854	8.9%
@kyliejenner	275	2.8%
@ladygaga	541	5.4%
@mrbeast	112	1.1%
@therock	726	7.6%
@willsmith	731	8.5%
@zachking	616	6.2%

Table 3. Comments in Final Dataset

	Comments Obtained from API	**Comments in Final Dataset** (duplicates and blanks removed)
Politicians		
@aocinthehouse	10,000	8,254
@bernie	1,332	335[1]
@Biggretchwhitmer	3,567	2,528
@corybooker	3,146	2,590
@gavinnewson	7,138	5,818
@ilhanmn	4,058	3,732
@jeffjacksonnc	10,000	9,817
@Johnfetterman	7,297	6,122
@Jon	8,345	8,283
@repbowman	2,822	2,371
TOTAL	**57,705**	**49,850**

Celebrities		
@addisonre	9,317	8,654
@billieelish	10,000	9,725
@charlidamelio	10,000	9,999
@jamescharles	9,636	8,782
@kyliejenner	9,780	9,505
@ladygaga	10,000	9,459
@mrbeast	9,941	9,829
@therock	9,598	8,872
@willsmith	8,564	7,833
@zachking	10,000	9,384
TOTAL	**96,835**	**92,042**

[1] The discrepancy between the comments in the dataset and the total number of comments on the videos (see Table 1) is likely a function of the API queries being run on a later date than when the comments were counted manually.

Analysis

To examine harassment within TikTok comments posted to public figure videos, we utilized two approaches. First, we examined a number of keyword lists available online that contained swear words, words with sexual meaning, racial slurs, and similar offensive terms. We eventually settled on the open-source List of Dirty, Naughty, Obscene, and Otherwise Bad Words (LDNOOBW) available at GitHub and used by Shutterstock, Slack, OpenStreetMap, and Google for content moderation and data sanitization purposes, as well as to train AI models. We believe this list of 403 keywords serves as an initial indicator of harassment and abuse. We identified the number of offensive keywords and phrases within each comment and, as a result, obtained one angle of the picture of how much harassment and abuse is received by the top politicians and celebrities.

The second phase of this project involved manual review of 28,686 comments (approximately 20%) from the corpuses of comments to the politicians' and celebrities' videos. Among the celebrities, 18,365 comments were hand-coded. For the politicians, 11,576 were hand-coded. Raters reviewed comments for evidence of general harassment, sexual harassment, hate speech, threats, or any other abuse directed toward the creator. Those which filled the "Other" category were then reviewed yet again afterward to identify if the comments they reflected could be grouped together in an intelligible way (e.g., perhaps another category would surface).

Identifying that an offensive keyword exists within a live comment on the platform may signal a clear instance of abuse. However, the keywords in the list may be used in innocuous ways, such as in banter, the expression of sarcasm, wit, or humor, to quote the words of someone else, for the purposes of emphasis, etc. We did not consider differences in opinion or logical arguments as harassment (e.g., @corybooker: I'm still mad at you for voting with big pharma and against me. I will never forget.).

Results

Keyword Analysis

As mentioned above, we programmatically examined the extent to which words and phrases from the LDNOOBW appeared in the dataset. This was done in a bag-of-words approach to represent and understand the contents of each comment. As displayed in Tables 4 and 5, only 59 of the 403 unique words and phrases from the list appeared among the politician comments, while 90 of the 403 words and phrases appeared in the celebrity dataset. Commonly used swear words appeared the most often among all of the creators reviewed. Among the politicians (Table 4), @gavinnewson, @jeffjacksonnc, @aocinthehouse, and @Jon had the most instances of these words and phrases, but this is a bit misleading since these creators also generally had the most comments in the dataset. It is difficult to standardize this measure since creators had different frequencies of comments, and comment word length varied considerably. For the celebrities (Table 5), @jamescharles had significantly more instances of the words and phrases. Even though comment word length also varied in this group, at least the total number of comments analyzed was relatively similar (ranging from 8,654 to 9,999).

Table 4. Keyword Analysis - Politicians

	Total		@aocinthehouse	@bernie	@Biggretchwhitmer	@corybooker	@gavinnewson	@ilhanmn	@jeffjacksonnc	@Johnfetterman	@Jon	@repbowman
shit	253		51	1	7	13	45	6	50	20	36	24
fuck	171		20	1	10	5	44	4	18	12	24	33
fucking	144		25	0	7	9	20	1	41	14	25	2
ass	134		17	0	8	4	36	10	15	12	22	10
bullshit	50		11	0	0	2	4	0	19	0	9	5
suck	46		6	0	4	4	10	0	9	6	5	2
sexy	38		2	0	0	0	0	0	0	2	34	0
sucks	37		4	0	2	1	3	1	9	8	4	5
sexual	32		22	0	2	1	3	0	1	0	3	0
butt	24		3	0	2	1	6	0	4	2	6	0
sex	24		2	0	0	0	10	2	8	2	0	0
rape	21		3	0	0	0	7	0	7	3	0	1
bitch	18		2	0	1	0	3	1	7	0	4	0
fuckin	15		0	0	3	0	1	1	4	2	3	1
🖕	11		3	0	0	1	4	0	0	1	0	2
piece of shit	10		4	1	0	0	4	0	0	0	0	1
god damn	9		1	0	2	1	0	1	1	1	2	0
viagra	9		0	0	0	0	0	0	9	0	0	0
asshole	8		2	0	0	1	1	0	1	2	0	1
shitty	8		1	0	1	0	1	1	1	1	2	0
dick	7		0	0	0	0	2	0	4	1	0	0
sexually	6		5	0	0	0	0	0	0	0	1	0
rapist	5		5	0	0	0	0	0	0	0	0	0
anus	4		0	0	0	0	0	1	0	0	3	0
bitches	4		0	0	1	0	1	0	0	0	2	0
incest	4		0	0	0	0	1	0	3	0	0	0
bastard	3		0	0	0	0	3	0	0	0	0	0
pedophile	3		0	0	0	0	2	0	1	0	0	0
pussy	3		1	0	1	0	1	0	0	0	0	0
sexuality	3		0	0	0	0	3	0	0	0	0	0

big black	2		0	0	0	0	0	0	0	0	0	2
hardcore	2		1	0	0	0	0	0	1	0	0	0
pissing	2		1	0	0	0	0	0	1	0	0	0
porn	2		2	0	0	0	0	0	0	0	0	0
pornography	2		0	0	0	0	2	0	0	0	0	0
raping	2		0	0	0	0	2	0	0	0	0	0
strap on	2		0	0	1	0	1	0	0	0	0	0
tits	2		2	0	0	0	0	0	0	0	0	0
twat	2		1	0	0	0	1	0	0	0	0	0
twink	2		0	0	0	0	0	0	0	0	2	0
anal	1		0	0	0	0	0	0	0	1	0	0
bestiality	1		0	0	0	0	0	0	0	0	1	0
domination	1		0	0	0	0	0	0	0	0	1	0
double penetration	1		0	0	0	0	0	0	1	0	0	0
fecal	1		0	0	0	0	1	0	0	0	0	0
genitals	1		0	0	0	0	0	0	1	0	0	0
girl on	1		0	0	0	0	1	0	0	0	0	0
humping	1		1	0	0	0	0	0	0	0	0	0
jack off	1		1	0	0	0	0	0	0	0	0	0
motherfucker	1		0	0	0	0	1	0	0	0	0	0
orgy	1		0	0	0	0	0	0	0	1	0	0
paedophile	1		0	0	0	0	1	0	0	0	0	0
playboy	1		1	0	0	0	0	0	0	0	0	0
poof	1		0	0	0	0	0	0	1	0	0	0
rectum	1		0	0	0	0	1	0	0	0	0	0
snatch	1		0	0	0	0	1	0	0	0	0	0
spic	1		1	0	0	0	0	0	0	0	0	0
vagina	1		1	0	0	0	0	0	0	0	0	0
whore	1		0	0	0	0	0	1	0	0	0	0
TOTAL	**1143**		**202**	**3**	**52**	**43**	**227**	**30**	**217**	**91**	**189**	**89**

Table 5. Keyword Analysis - Celebrities

	Total		@charliedamelio	@addisonre	@mrbeast	@zachking	@willsmith	@therock	@ladygaga	@billieelish	@kyliejenner	@jamescharles
shit	102		4	4	3	4	13	9	13	25	6	21
fucking	93		5	8	2	2	13	5	10	17	13	18
butt	83		3	2	2	5	6	1	1	3	21	39
fuck	67		1	6	5	2	13	7	5	10	3	15
ass	60		2	4	6	2	17	5	3	2	4	15
xx	43		4	9	0	0	0	1	5	5	3	16
bitch	21		0	4	0	0	5	0	4	0	1	7
xxx	17		1	3	0	0	1	1	3	2	3	3
suck	15		0	1	0	0	4	5	0	1	0	4
sucks	11		1	0	1	0	1	1	1	3	0	3
sexy	9		0	2	0	0	1	2	2	1	1	0
fuckin	8		0	0	0	0	1	0	0	1	1	5
girl on	8		1	2	0	1	0	0	1	3	0	0
god damn	8		1	0	2	0	1	0	1	2	0	1
negro	5		0	0	0	3	2	0	0	0	0	0
bitches	4		0	0	0	0	0	0	0	3	0	1
cunt	3		0	1	0	0	0	1	1	0	0	0
sex	4		0	0	0	0	1	2	1	0	0	0
sexual	4		0	1	0	0	0	0	0	0	2	1
sexuality	4		0	0	0	0	0	0	0	4	0	0
shitty	4		0	2	0	0	1	0	0	1	0	0
asshole	3		0	0	0	0	0	1	1	0	0	1
boob	2		0	0	0	0	0	0	0	0	1	1
bullshit	2		0	0	0	1	0	1	0	0	0	0
taste my	2		0	0	1	0	0	0	0	0	0	1
twink	2		0	0	0	0	0	0	1	0	0	1
anus	1		0	0	0	0	1	0	0	0	0	0
boobs	1		0	0	0	1	0	0	0	0	0	0
cock	1		0	0	0	0	0	1	0	0	0	0
homoerotic	1		0	0	0	0	0	0	0	0	1	0
hooker	1		0	0	0	0	0	0	1	0	0	0
horny	1		0	1	0	0	0	0	0	0	0	0

make me come	1		0	0	0	0	1	0	0	0	0	0
masturbation	1		0	0	1	0	0	0	0	0	0	0
mong	1		0	0	0	0	0	0	0	1	0	0
nude	1		0	0	0	0	0	0	0	0	1	0
panties	1		1	0	0	0	0	0	0	0	0	0
piece of shit	1		0	0	0	0	1	0	0	0	0	0
pissing	1		0	0	0	0	0	0	0	0	1	0
pussy	1		0	0	0	0	0	1	0	0	0	0
rape	1		0	0	0	0	0	0	1	0	0	0
snatch	1		0	0	0	0	0	0	1	0	0	0
swinger	1		0	0	0	0	1	0	0	0	0	0
vagina	1		0	0	0	0	0	0	0	0	0	1
🖕	1		0	0	0	0	0	0	0	0	1	0
TOTAL	**603**		**24**	**50**	**23**	**21**	**84**	**44**	**56**	**84**	**63**	**154**

Keyword comparison approaches can provide a general sense of toxicity within social media comments, but are not without limitations when attempting to automate text moderation. Chiefly, keyword lists are not adaptive and often fail to include a variety of intentional misspellings, abbreviations, slang, novel vernacular, or other encapsulations of the same harassing sentiments that still can cause harm.

For example, the following words and phrases were not highlighted in our keyword search, despite their clear intent:

- Tuck Frump 💩 (@aocinthehouse)
- Are you f*cking kidding me? 🤬🤬🤬 (@aocinthehouse)
- sounds like the American people are getting f***ed again bipartisan just means double penetration (@jeffjacksonnc)
- she's having fun on tik tok bitchessss (@kyliejenner)
- she's your worst nightmare you MOTHERF*CKERS (@addisonre)

The bag-of-words model is also limited in that an exhaustive list is simply untenable, given all the possible permutations of proscribed words. Not only that, many times instances of "naughty" words are perfectly legitimate or even complimentary, such as this comment to @billieelish: "What the fuckkk this is beautiful." Take this example from a comment to @aocinthehouse's video: "u keep putting in people who just got their GED & we will put in people who graduated Magna Cum Laude with a double major. that's repub vs dem today." The word "cum" is on the LDNOOBW and yet in this instance it is not used in a naughty way. Similarly, misspellings can get caught up in the keyword search. This is a comment on @therock's video: "Jack my favorite movie is the school of rock. couldn't cunt how menne times iva see it."

Furthermore, nuance is lacking in that intent, motive, context, humorous or metaphorical use, and cultural appropriateness cannot be surmised through basic keyword comparisons. As such, it is likely that false positives and false negatives will surface, muddying the proverbial waters when it comes to an understanding of the extent of abuse in TikTok comments. To address this concern and given that content moderation often involves ambiguities and complexities in interpretation, we engaged in human review to account for the possibility of the misidentification or misclassification of problematic comments.

Manual Analysis

Overall, the research team manually reviewed nearly 29,000 comments (10,281 from the politician videos and 18,365 from the celebrity videos) to search for instances of harassment, sexual harassment, and other forms of abuse (see Table 6).

Table 6. Manual Analysis

	Total Comments	Number (%) Manually Reviewed
Politicians	49,850	10,281 (20.6%)
Celebrities	92,048	18,365 (20.0%)

We identified 472 comments that could be considered general harassment (see Table 7). There were more instances of harassment noted in the data from the politicians (2.1%) than the data from the celebrities (1.2%), though overall, these were a small proportion of the total data (only about 1.6% of the comments in total). Table 8 displays examples of harassment observed for each user.

We observed fewer instances of sexual harassment within the comments (less than 1%). There were a number of comments directed toward the creator’s appearance, both among the politicians (shes a beautiful lady [@aocinthehouse]) and the celebrities (U STILL HOT ASFFFF THO [@billieelish]). These could be considered sexual harassment if we knew more about the intent of the user or the impact on the creator. Importantly, most definitions of sexual harassment include an element of undesirability. For example, U.S. law defines sexual harassment as “*unwelcome* advances, requests for sexual favors, and other verbal or physical conduct of a sexual nature” (emphasis added).[5] As such, it is difficult if not impossible to categorize many of the appearance comments as sexual harassment without asking the creator if the comments are unwanted.

And while many of these comments were directed at female creators, the most sexually inappropriate comments observed (by a wide margin) appeared on Jon Ossoff's (@Jon) account:

"you're so hot dad"
"I was busy thinkin bout being your Monica Lewinsky 😳
"WHY ARE U SO FINE"
"And zero days in me 😔"
"i don't even care who sees this, jon you are fine as hell and can 100% solar my panels any day"
"breed me"
"Jon Ossoff is the hottest politician ever omg"
"yes dad *bends over*"
"Sorry you have to fill out a form before coming"
"MARRY ME JON I LOVE YOU, I MOVED TO GEORGIA JUST TO BE UNDER UR CONTROL DADDY"

In fact, there were dozens of references to @Jon being "hot." Again, it is impossible for us to determine if these comments are accurately defined as sexual harassment, but the fact that they remain on his account (many months after being posted) might give an indication that he is not bothered by them. It is reasonable to assume that creators who do not want these types of comments on their videos remove them and probably block the users who posted them.

While we did not observe a significant amount of comments that included hate based on race or gender identity, there were a few isolated examples. For instance, there were several comments directed toward Ilhan Omar in response to her video about a visit to Somalia that could be interpreted as racist/hate speech ("Please don't come back" "Blah blah blah. Why don't you stay there then." "she looks good only there 😁😁"). There were two comments directed toward @aocinthehouse that referenced her race, one more direct than the other: "Hey there spic" "Would you stop eating all those burritos? It's noticeable now."

Table 7. Instances of Harassment Observed in Comments

(% of comments reviewed for that creator)

	Number of Comments Analyzed	Harassment	Sexual Harassment	Total
Politicians				
@aocinthehouse	1,749	35 (1.7%)	2 (0.1%)	37 (1.8%)
@bernie	71	2 (0.7%)	0 (0.0%)	2 (0.7%)
@Biggretchwhitmer	591	11 (1.5%)	0 (0.0%)	11 (1.5%)
@corybooker	556	6 (0.9%)	0 (0.0%)	6 (0.9%)
@gavinnewson	1,195	39 (2.8%)	0 (0.0%)	39 (2.8%)
@ilhanmn	717	61 (8.0%)	0 (0.0%)	61 (8.0%)
@jeffjacksonnc	2,020	19 (0.9%)	0 (0.0%)	19 (0.9%)
@Johnfetterman	1,267	57 (3.9%)	0 (0.0%)	57 (3.9%)
@Jon	1,655	14 (0.8%)	70 (4.2%)	84 (5.0%)
@repbowman	460	1 (0.2%)	0 (0.0%)	1 (0.2%)
TOTAL	**10,281**	**245 (2.1%)**	**72 (0.6%)**	**317 (2.7%)**

Celebrities				
@addisonre	1,731	38 (2.0%)	4 (0.2%)	42 (2.2%)
@billieelish	2,017	13 (0.6%)	0 (0.0%)	13 (0.6%)
@charlidamelio	2,030	12 (0.6%)	0 (0.0%)	12 (0.6%)
@jamescharles	1,706	53 (2.8%)	0 (0.0%)	53 (2.8%)
@kyliejenner	1,880	27 (1.4%)	1 (0.1%)	28 (1.5%)
@ladygaga	1,890	17 (0.9%)	0 (0.0%)	17 (0.9%)
@mrbeast	1,975	6 (0.3%)	1 (0.1%)	7 (0.4%)
@therock	1,747	11 (0.6%)	0 (0.0%)	11 (0.6%)
@willsmith	1,560	48 (2.8%)	0 (0.0%)	48 (2.8%)
@zachking	1,829	2 (0.1%)	0 (0.0%)	2 (0.1%)
TOTAL	**18,365**	**227 (1.2%)**	**6 (< 0.1%)**	**235 (1.2%)**

Table 8. Examples of Harassment Observed in Comments

Politicians	
@aocinthehouse	“Wow this woman is a complete idiot and if I asked for a beer she’d bring me water. What a piece of shit how did she get voted in” “AOC LOOKS LIKE SHES BEEN EATING WELL 😆” “Your a JOKE!!!!! A child in an adult position! Do something to help Americans and earn your paycheck!!!! He will be back!”
@bernie	“SHUT UP! You're creepy!😂😂😂”
@Biggretchwhitmer	“This is why whitmer is brain dead” “She’s a good liar. They conspire to strip people of their rights. This is the worst state ever” “No one wants to hear your virtue signaling nonsense”
@corybooker	“DRAMA QUEEN! Just because you talk slowly doesn't validate any of the bs from your mouth!” “Don’t like Cory. Partisian hack. Making it political and won’t make hard decisions.”
@gavinnewson	“what a tool! virtue signalling” “Lame. How about you fix ca problems? Shootings hourly. Guns everywhere!! Crime everywhere and you do nothing!!! You should be ashamed!!!” “Get lost fix California! Democrat scum!”
@ilhanmn	“Age is catching up with you dhilo” “you are a terrible person.u move here and succeed because of this country.Then u have the nerve to disrespect the same country that helped you.pos” “What a great day for America us. Omar gets the boot. Bye Bye you RACIST 👏🏽” “Good thing is this life is short but your soul will burn for eternity in Hell”
@jeffjacksonnc	“You’ve done a great job of slowly ramping up the partisanship with this account… I don’t blame you just saying what I see.” “Ignore him, he is a puppet !” “This is just insanely stupid. 😂😂”
@Johnfetterman	“What an embarrassment to our state.” “Bro, at least oz can put together a proper sentence, since his stroke fetterman can’t even speak properly” “I loved you in goonies!” “I didn't know we were hosting the special Olympics”
@Jon	“you dems assured me and my son are homeless i have 0 respect for you're fucking games dont ever expect my vote” “what a fuckin joke you are”
@repbowman	“Talking to you would be less exciting and productive than watching paint dry. At least with the paint I have the satisfaction of accomplishment.”
Celebrities	
@addisonre	“Not another movie Addison… YOU CANT ACT” “NO NOT ADDISON RAE WITH ANOTHER SHITTY MOVIE”
@billieelish	“I aspire to sing like this with no tune” “Lost all respect for her, same”
@charlidamelio	“She is so spoiled literally crying cause she lost hundred million followers but me with my 1 follower like what” “I have no idea how she still gets over 100 views 💀”
@jamescharles	“I think James is missing a few brain cells. Of course the ombré is gonna go away if you blend it back and forth. The girl in the vid didn’t and got it”
@kyliejenner	“Stop being miserable. Touch some fucking grass weirdo”
@ladygaga	“She doesn’t look like Britney at all. She just looks like she’s on Ozempic and got her fillers on” “Im guessing eye lift, buccal fat removal, lip filler, cheek filler idk but this industry is tough,I hope she got it done for herself not others :(“
@mrbeast	“I used to like mr beast but this is the kinda stuff Caden boof would do 😂” “Excruciating watch”
@therock	“please stop ruining movies” “I remember when the Rock first popped on the scene. He was slim @ had hair. He looked like a Latino pimp. Now he looks black & sounds like OJ Simpson.”
@willsmith	“You still come out like nothing happened? shame on you.” “Dude get out of our Africa. We don't want your garbage influence”
@zachking	“U look getting old buddy.” “you geting old my friend”

At times it was difficult to determine if a negative comment was directed at the creator, someone the creator references in their video, or another commenter. For example, there were a number of instances of comments such as “fuck that shit” or “fuck that guy.” Another common example is exemplified in this comment: “Who the fuck are you and who cares” (@therock). If this was directed at @therock, then it could be considered harassment. If it was directed at another user’s comment, then it would not be (at least not toward the creator). We only coded comments as harassment if it was reasonable to conclude the comment was directed at the creator. Of course, harassment occurring among and between commenters in the comment threads is worth studying, but without knowing who specifically a comment was directed toward, this becomes a difficult task. It would help if the API indicated if the comment was a reply to the original content or to another comment. This could be done by inserting the username of the person the reply was direct to within the comment (e.g. @fakename what are you talking about????)

Related to this, it was sometimes difficult to know whether a comment was harassing in nature without knowing the content of the video. On a number of occasions, our human review team felt the need to review the video in order to learn more about what was being discussed. This is a reasonable approach when only reviewing comments to 20 videos, but becomes challenging at scale.

API Limitations

While using TikTok's Research API, we experienced a significant number of issues which delayed progress. We share those issues below so that they can be addressed by TikTok developers before the Research API is opened to researchers globally.

Sorting

When querying for TikTok video data, there are no provided parameters to facilitate sorting by certain fields. We are not able to get the latest 10 videos based on a particular username, for instance. The API only provides the videos in random order, or not in random order (using the *is_random* parameter). If one retrieves the videos not in random order, the documentation doesn't make clear how they are selected. Is it by most recent? Most liked? Interestingly, the Legacy API (according to the documentation) does allow for queries that provide videos "sorted by create_time in descending order." It is not clear why the new Research API does not provide this functionality.

Comments are sorted and provided only by most-liked. It is our recommendation that the API provide the ability to retrieve comments sorted by other parameters such as random, most recent (reverse chronological order), most replied-to, most influential (e.g., those who have been verified, those who have the most followers), etc.

Timeframes

Currently, the API restricts the ability of researchers to obtain more than 30 days of video data when querying by username. To overcome this, a researcher can simply make the same query 30 days by 30 days. However, this seems inefficient. If the API can provide, for example, one year's worth of video data, why force the researcher to retrieve it in 30-day increments? Admittedly, this may be set to lessen server load, but does introduce yet another obstacle to overcome.

The "created_date" field may be redundant. A researcher can simply restrict the start date and end date to a specific day, and that provides the created date. Perhaps the documentation can better explain how it might be used; indeed, providing more short, simple explanations and examples in the API documentation would be welcomed.

Usernames

The username of the video creator, and not the commenter, is provided by the Comments API. If the Comments API is providing one, why not provide the other? The documentation indicates that TikTok removes personally identifiable information from comment text. Is this the reason why the username of the commenter is not provided? Explaining this in the documentation can help researchers accomplish their goals more efficiently.

Querying for usernames was very problematic. It is not clear whether ampersands, quotation marks, or other characters should be used around usernames. We also used the EQ condition operation for querying usernames according to the API specifications. This was supposed to give us exact matches for the username. Even still, querying for some returns results for multiple permutations of that username. By way of example, querying for “@jon” *also* returned video data from other users:

jon..jafrii05
jon.286
............jon
jon.706
..........jon
jon.1998_
._.jon

Similarly, when querying for “bernie” video data was *also* pulled from:

bernie.1990
bernie.613
alps2.bernie
alps2.bernie

It is not clear how to retrieve video data solely from, for example, “@bernie” - and no other usernames in which the character string “bernie” also occurs - even when carefully following API documentation.

Rate Limits

In the Research API documentation, there are no details on rate limits. It is reasonable to expect that this would be provided. The documentation for Server API v2 (not the Research API) indicates that the rate limit is 600 per sliding minute window, and a researcher is left to assume that it is the same for the Research API.

It is also not clear what the daily limits are. Multiple times, we reached our daily limit on API calls and had to delay progress on the project until the next day. While having limits in place is necessary, it is arguable that they are set too low for any meaningful data collection. It is also not clear when daily limits reset. Is it at 12am GMT? For time-critical projects, every hour matters and so this should be detailed in the documentation.

Ultimately, we settled on a .2 second delay between API calls to fetch comments, exclusive of the time taken for the API request, response, file writes, data processing, etc. As such, the .2 second delay ensured that 5 requests were sent per second at most, and so our script was limited to perform 300 requests per minute. We hit our limit at around 1,000-1,050 calls per day.

When we reached out to the Research API Support Team for clarity on rate limits, it took 9 days to receive a response. This was problematic for a project like ours with such a short window (only about 4 weeks). Hopefully the Support Team will have enough resources to response to questions more quickly moving forward. Ultimately the response that we received stated that the restrictions are "100 per minute as the rate limit and 100 per day for the quota limit" which in fact did not align with our own experience using the API. The documentation should be clear about what the limits are.

The documentation states that a researcher should send a message to Research API Support asking for a rate limit increase. We reached out on June 13th, 2023, but never heard back. As such, we had to make do and elongate the data fetching aspect of the project across multiple days to get what was needed, which put us in a time crunch to complete the final report by June 30, 2023.

Comments

Depending on the creator, video comment data often includes empty content in the text field. The comment is returned with an ID, a video ID, a creator ID, timestamp, number of likes, number of replies, etc., which seems to indicate that something was there. Was the text of the comment deleted by TikTok? Was it hidden internally by TikTok before being fetched? Was it deleted by the video creator? Was it deleted by the user who made the comment? Was the user's account deleted and all comments associated with that user removed? When looking at this project through the lens of Trust and Safety, knowing what the deleted/redacted comment was and/or the reason(s) behind its removal would greatly inform our understanding of the extent of aggressive, harassing, or otherwise problematic comments on TikTok.

In comments, there is no "hashtag" field like there is for the video data. Is this because a creator adds hashtags separately from the video description text, while the commenter adds it manually within their comment (and as such it is plain text)?

Since both, though, are interpreted in the same way programmatically, shouldn’t commenter hashtags also be a field provided? To be sure, manual searches across the data can still occur, and commenters use hashtags much less frequently than creators. But if a specific hashtag is trending in comments, or used as a vector for harassment en masse, it seems like it would be easy to provide it as a separate field.

Other Observations

Since comments are liked in real time, and the API only returns comments sorted by most liked, a second request for comments from a username or a video may not return the same set of comments. This is to be expected, but is worthy of note.

What time zone should date parameters be set to? GMT? This is not clear and the documentation should be detailed enough so that researchers have this information.

Limitations to the Current Analysis

The largest barrier with the current project was the allotted time we had to complete it. We only had about four weeks from the time we were granted access to the Research API to when our report was due. Given more time, it would be easy to expand the current pilot study by exploring more creators, more videos, and more comments. It also would be easy to search hundreds of thousands of comments for the presence of specific keywords, and a larger subsample of comments could be manually reviewed with more time and resources. Additionally, the time constraints precluded our ability to use multiple raters to review the same comments so as to provide a measure of interrater reliability. Future qualitative research using the API should do so to ensure that interpretations of the sentiment of each content are consistent and unbiased.

It also warrants mention that without specifically asking public figures as to whether they have been harmed by content in the comments (or private messages) they receive, researchers only have a partial picture of the extent of harassment they experience on TikTok. In attempting to understand their experience when compared to private figures, public figures simply may face different types, frequencies, and intensities of interaction on TikTok. This, of course, remains to be determined and leads us into the next section of this report.

Directions for Future Research

TikTok may very well provide to all creators (public figures and private figures alike) a satisfactory array of safety features within the platform to control and moderate the engagement of others on their videos. They may also be appropriately preventing harassing and abusive comments through their automated and manual moderation solutions. However, intentionally soliciting and learning from additional stories and experiences of users through further research would likely germinate or illuminate new in-app, educational, or policy-based initiatives that can support and protect them. It seems valuable, therefore, to supplement findings from the current examination with richer and more nuanced perspectives from those on the receiving end of these comments. Doing so may uncover, for example, more personalized or coordinated attacks via private message instead of on publicly viewable comment threads.

Specific to the Research API, future inquiries might explore positivity within comments. We observed many examples of supportive comments: (“MR BEAST! You changed my life! You helped me and others to put good into the world. LOVE YOU MR BEAST!!” [@mrbeast]; “Thank you so much for this. It honestly has helped me because I’m in a bad funk.” [@therock]; “Thank you so much for being in our/my life ❤️you helped me through so much, can not imagine a world without your beautiful soul anymore ❤️❤️” [@ladygaga]; “Thank you for your videos. They have really helped me understand how government really operates.” [@jeffjacksonnc]).

We also saw a number of comments where users stepped up to defend the creator or at least attempted to maintain a civil comment section: (“People need to stop commenting this, it's not ok. It's not your place to speculate or comment on someone's weight” [@ladygaga]; “y’all were all free brit movement’’ but at the same time you’re being toxic towards Gaga the same way the media were with Brit. just leave her alone🤯” [@ladygaga]; “Can we stop hating on her she’s actually sweet” [@charlidamelio]; “stop hating will smith” [@willsmith]).

Instead of relying on large lists of proscribed words, future research would do well to search for more targeted keywords. For example, while it is unlikely that scholars interested in self-harm will identify content by searching for the standard words or hashtags associated with these behaviors (e.g., #cutting, #selfharm, #selfharmmm, #hatemyself, #selfharmrecovery and #selfharmawareness)—because TikTok already prohibits some or all of these words—researchers in the trenches can search for terms young people are using at the current moment to circumvent automated moderation. Given the real-time nature of the API, this tool could be nimble enough to examine words as they change monthly, weekly, or even daily. (Notably, we did not identify any comments that appeared to indicate any reference to self-harm.)

The Research API could also be used to assess the real-time actual extent of purported viral trends on the app. For example, if conventional wisdom suggests a particular video, argument, or dance is “going viral” on TikTok, a quick search using the API could determine the actual magnitude of videos with that type of content on the app. We are reminded of the Momo craze on YouTube and other social media platforms in 2018 and 2019 where fear spread among parents about “Momo” demanding their children engage in dangerous tasks. When concerns like this arise, the API could be utilized to simply assess how many times “Momo” is mentioned, and additional investigation into these instances could be carried out by researchers to understand the context, the ways in which the information is being spread, the extent of its reach, the amount of engagement it is receiving, etc. Moreover, researchers could study viral incidents like this after the fact to ascertain their spread and consequences.

Finally, the bag-of-words approach to automated content moderation should also be supplemented in future research endeavors with natural language processing techniques such as weighting schemes to determine how important a word is within a comment via its frequency and rarity, dimensionality reduction to represent the most important words after analyzing patterns or themes, or word embeddings to assign similar numbers to terms that are related to each other for faster analysis and automatic flagging.

Summary

The current project had two objectives: (1) Explore the TikTok Research API and identify possible limitations to researchers who desire to use it in their scholarly initiatives; and (2) Examine the nature and extent of harassment within comments to videos posted by public figures.

The Research API is a powerful tool that allows for the retrieval of public TikTok account data including user profile details, video captions, comments, as well as number of likes, comments, and more. We encountered a number of challenges when using the tool, but hope that TikTok will be able to remedy some of the limitations we identified and/or clarify their documentation regarding the issues we confronted.

The data we were able to collect using the Research API shed some light on the nature of harassment in comments to the TikTok videos of the public figures we explored. Overall, our inquiry identified relatively little harassment within the comments of TikTok videos. Only about 1% of the comments reviewed included any type of harassment, and the vast majority of sexual harassment was on one account (@Jon). We identified only a few isolated examples of what may be considered hate speech, and we did not identify any instances of threats.

There are two noteworthy caveats in the conclusions that can be drawn from this analysis, however. First, we were not able to analyze all of the comments to a creator's video. We were limited to a maximum of 1,000 comments per video, and when there were more than 1,000 comments made to a particular video, we were given the most liked comments. In order to do this analysis properly, we either need to have access to all of the comments, or a random sample of all of them. Second, we need to know what is contained in the blank comments. If these are comments that were removed (either by the creator or by TikTok), then it is probable that they included content pertinent to this analysis. The real key in any analysis of TikTok comments is knowing that researchers have access to the full corpus of the comments. Otherwise, there remains too many questions about what was actually reviewed.

As referenced earlier, some public figure accounts (e.g., @bernie) have a surprisingly low number of comments and engagement when compared to the number of followers they have. It seems evident that a great deal of moderation is being done to keep clean the comment threads associated with his videos. For @jon and the disproportionate but relatively low number of comments he received that could be classified as sexual harassment, it seems that he or his team are not bothered enough to moderate their presence under his posted videos.

Solely focusing on the relative infrequent amount of general harassment, sexual harassment, hate speech, threats, and other forms of abuse found in this pilot project when considering the experiences of politicians and celebrities on TikTok, two general conclusions can be made. The default TikTok moderation approach seems to be keeping the worst forms of toxicity off the platform. Second, it appears that creators have adequate tools to control the display of certain comments under the videos they post, and at least some seem to be using them. Though the current analysis was constrained by a number of factors including project deadlines, API limitations, and the ever-present contextual caveats that exist when subjectively evaluating harassment and abuse, we are optimistic about the possibilities of the Research API to answer questions related to user experiences on TikTok.

References

1. Lee C. TikTok now has 150 million active users in the U.S., CEO to tell Congress. NBCNews.com, March 19, 2023. Accessed June 28, 2023. https://www.nbcnews.com/politics/congress/tiktok-now-150-million-active-users-us-ceo-tell-congress-rcna75607

2. TikTok. Research API Overview. TikTok.com. Accessed June 19, 2023. https://www.tiktok.com/transparency/en-us/research-api/

3. Vincent J. TikTok bans deepfakes of nonpublic figures and fake endorsements in rule refresh. The Verge, March 21, 2023. Accessed June 18, 2023. https://www.theverge.com/2023/3/21/23648099/tiktok-content-moderation-rules-deepfakes-ai

4. Suciu P. Do Social Media Companies Have The Right To Silence The Masses – And Is This Censoring The Government? Forbes, January 11, 2021. Accessed June 14, 2023. https://www.forbes.com/sites/petersuciu/2021/01/11/do-social-media-companies-have-the-right-to-silence-the-masses--and-is-this-censoring-the-government/?sh=6a2f538a48e2

5. Legal Information Institute. 29 CFR § 1604.11 - Sexual harassment. Cornell Law School. Accessed June 20, 2023. https://www.law.cornell.edu/cfr/text/29/1604.11