Suggestive Scrolling:

An exploration into the effectiveness of TikTok Moderation Settings regarding Sexually Suggestive Content

April 24, 2024

Word Count:

4,741

Abstract

This research aims to test the effectiveness of TikTok's moderation of sexually suggestive content, through its opt-in moderation settings offered to users. Each account, one without and the other two with different opt-in moderation settings viewed 100 videos under 3 hashtags that were deemed to have sexually suggestive content and coded for frequencies of auditory and visual sexually suggestive content. The frequencies were then compared using a two-factor ANOVA test, and it was determined that there was not a significant difference in the amount of codes shown based on the account type. This implies that the moderation settings were not effective at filtering sexually suggestive content in this scenario and exposed the user to the same amount of sexually suggestive content regardless of the moderation settings turned on.

Literature Review

During early adolescence, which is often considered to be ages ten to twelve, children are most easily influenced by the content they consume in the media. This influence continues for the remainder of their adolescence, but the impact the content has on their thoughts and behaviors decreases after the age of twelve, notably if the influence could lead to consequences the adolescent knows is bad if acted upon (Braams et al., 2018). A major way children consume media in the current day is through websites and online applications, mainly social media platforms (Evli et al., 2023). This online influence continues to increase as personal smartphone ownership among adolescents increases, with the Pew Research Center reporting that 95% of teenagers aged 13 to 17 own a smartphone and 46% of teenagers are on the internet constantly in 2023, almost double from the 24% reporting such in 2015 (Anderson et al. 2023).

As adolescents' access to the internet becomes easier and more widespread, they are exposed to more content that could potentially influence their thoughts and behaviors (Te'eni-Harari et al., 2007). Although adolescents can use the internet to quickly access information and contact friends and family, certain types of information and content found on the internet, such as violent and sexual content, can have negative effects when consumed (Kharlamova & Erofeeva, 2022). Sexual content is found on both the internet and social media platforms adolescents commonly use, such as Snapchat, Instagram, and TikTok (Lewis et al., 2018). Sexual content can be in the forms of explicit sexual content, as is content that directly shows nudity or sexual acts, and sexually suggestive content, which is content made up of writing, visuals or audio content with evident sexual undertones (Paasonen et al., 2019). This sexual content is widespread, with 54% of teenagers under the age of 13 having seen online pornography, increasing to 73% by the age of 17. Out of the adolescents who were exposed to such content, over half of them reported seeing it accidentally through social media (Robb & Mann, 2022). Exposure to explicit sexual content increases the likelihood of viewing suggestive content, as adolescents may purposefully search for sexual content on social media after being exposed to it (Setyawati et al., 2020). As suggestive sexual content is more frequent on social media platforms than explicit sexual content, they will likely find and be exposed to suggestive content while looking for explicit content (Paasonen et al., 2019).

Visual exposure to sexually suggestive content during adolescence, which often features sexual poses and facial expressions, is associated with an increase in sexual characterization of themselves and others, along with engaging in sexual activities at a younger age and more frequently than those not exposed to the content (Coyne et al., 2018). An auditory exposure to such content, mainly through listening to sexually explicit song lyrics, has also been associated

with adolescents beginning to engage in romantic and sexual relations at a younger age than those who do not listen to these types of songs (Wright & Qureshi, 2015). If this sexual content is mainly made up of sexual stereotypes and objectification of women, it is also associated with higher levels of one sexually objectifying themself, though this effect is most prevalent in women (Karsay & Matthes, 2016). There has been a notable increase in the number of popular songs containing mentions of sexual content, with a jump from 18% of songs in the U.S. Billboard *Top 40 Single List* containing sexual references in 1960 to 41.7% in the 2000s, possibly exposing adolescents to this type of content more frequently (Christenson et al., 2018). As songs are often used alongside social media posts, this exposure to sexual content through the music accompanying posts is possible online as well (Wright & Rubin, 2016). To lower the number of sexual references in popular music while still allowing children to listen to it, music censorship is often used. The censorship of sexual content, mainly found in songs through the use of 'bleeping' or cutting out certain words or phrases, appears to have the same effects as listening to the uncensored song, with listeners having their thoughts and behaviors influenced about equally by the music (Wright et al., 2018).

One consequence of adolescents having their thoughts and behaviors influenced by sexually suggestive media is beginning to engage in sexual activities at a younger age. This is tied to increased rates of contracting and transmitting sexually transmitted diseases, and unwanted pregnancies due to a lack of knowledge about safe sex practices (Coker et al., 1994). The effects of this go beyond physical, with adolescents feeling ashamed, guilty and possibly experiencing anxiety and depression when engaging in this accelerated sexual behavior (Wesche et al., 2017). Other researchers, such as Lebedíková et al., argue that exposure to explicit and suggestive sexual content is positive, as adolescents have the opportunity to learn about safe sex

practices and sexual health through this content. Despite the intention of the content not being sexually educative in nature, the researchers predict that this exposure could cause adolescents to put more thought into sexual behaviors they perform or seek out sexually educative content, overall benefiting them and possibly preventing unintended outcomes of accelerated sexual behavior (Lebedíková et al., 2022).

According to the Pew Research Center, 63% of teenagers reported using the social media platform TikTok at least once in 2023, with 17% of them using the app "almost constantly," and 32% using it multiple times a day (Anderson et al., 2023). TikTok is a social media platform that is made up of short videos created by users, often featuring audio of popular music (TikTok 2023). The use of TikTok in adolescents is mainly driven by factors such as building and maintaining relationships, escapism, and surveillance of others (Bucknell & Kottasz, 2020). Adolescents also find the algorithm TikTok uses to recommend content to be much more perceptive of their preferences and recommends more content they enjoy compared to other social media platforms, such as Instagram or Twitter, further motivating them to use this app (Taylor & Choi, 2022). This large amount of time spent on the app exposes users to many different types of content, such as comedic, educational, or informative (Siswanto et al., 2022). Some of this content can feature sexually suggestive themes, as a study performed by Soriano-Ayala et al. found that 25% of videos collected from accounts most followed by Spanish youth contained explicit song lyrics, and another 25% containing sexually suggestive visuals (Soriano-Ayala et al., 2022). These findings show the existence of this content on the platform despite them only looking at accounts based in Spain. There is also sexually educative content on the platform, content that focuses on raising awareness of safe sexual practices and

sexual health, which researchers predict causes adolescents to become more invested in their sexual health when being exposed to this content (Stoddard et al. 2023).

Explicit and sexually suggestive content are prohibited on TikTok, but explicit music lyrics and sexually educative content are not ("Sensitive and Mature Themes", 2023). To moderate this content and other content that is not allowed on the app, TikTok uses a mix of human moderation and moderation done by computer algorithms that visually detects prohibited content on the app ("Our approach to content moderation", 2023). Moderating sexual content in a video format on social media platforms has posed many challenges, with platforms such as Facebook, Instagram, and Twitter struggling to do so because of algorithmic moderation miscategorizing sexual content as other types of content that are allowed on the platform, and human moderation skimming over the content and not taking an appropriate amount of time to examine the content due to their high workloads (Peters, 2020). Moderating between sexually suggestive content and sexual education content has also posed a challenge, with content moderation algorithms often miscategorizing educative content as explicit or suggestive sexual content and taking it down (Young, 2021).

Approximately 14% of removed content from TikTok from July to September 2023 was sexual content, including sexually suggestive content ("Community guidelines enforcement report", 2023). Users are also given opportunities to use tools on the app to further moderate the content the app recommends to them, such as "Restricted Mode," a setting that automatically filters content to prevent the user from seeing what TikTok deems as mature, specifically content with drug or alcohol use, violence or sexually suggestive elements. There is also "Keyword Search," which lets the user enter specific words that they do not wish to see, and TikTok will filter the videos to ensure hashtags with these words will not be shown to the user.

TikTok largely advertises these moderation methods to the parents of users, since it allows them to have more control over the type of content shown to their children ("User Safety"). Despite these moderation measures being available, users have also found ways to intentionally post content without it being taken down by TikTok, utilizing methods such as algospeak, which replaces letters in words with numbers or symbols, to post about sexual content on the app (Steen & Klug 2023).

The existence of sexually suggestive content that goes against TikTok's Community Guidelines, and the methods users have to circumvent this moderation, along with the lack of research on this subject calls into question the effectiveness of TikTok's content moderation. Thus, a directed content analysis is needed to see if there is a significant difference in the amount of sexually suggestive content shown and the effectiveness of these opt-in content moderation policies. Directed content analyses have been used on TikTok in multiple studies to see if videos under certain hashtags or user profiles meet a certain set of criteria. Soriano-Ayala et al. performed a directed content analysis of sexually suggestive content on TikTok, using preset codes from other researchers to determine the frequency of certain sexually suggestive elements under user profiles (Soriano-Ayala et al., 2022). This study will take inspiration from them, using a directed content analysis to determine the frequency of sexually suggestive codes, thus the amount of sexually suggestive content shown to the user.

Methods

This study explores the effectiveness of opt-in moderation settings on TikTok at moderating sexually suggestive content. These settings were tested in environments where sexually suggestive content was expected, to see if these filters were effective at blocking this type of content, and if certain filters were more effective. Parents are advised by TikTok to use

these settings to regulate the type of content their children see and block inappropriate content, such as sexually suggestive content, highlighting the importance of this study to show if adolescents are being exposed to sexually suggestive content while having these settings active.

The data collected and analyzed for this study were 100 TikToks under three hashtags from three separate accounts, for a total of 900 TikToks. Hashtags, as defined by Merriam-Webster, are a way to digitally identify content about a specific subject, which for this study will be sexually suggestive content (Merriam-Webster). On social media hashtags are often used by creators to group content they post into categories, allowing other users to easily find content regarding certain topics (Ayu et al., 2022). This use of hashtags is also seen on TikTok, as it has options to let users add them to their videos and search for videos under specific hashtags.

Originally, to determine these hashtags, 100 TikTok videos were watched from a new TikTok account and then coded for any suggestive sexual themes with the same codes that are used to code the videos under the hashtags. Once this method was performed, it was evident that this would not be effective as the videos did not feature sexually suggestive content often enough to collect hashtags that were directly linked to sexual content. This caused a shift to intentionally look for hashtags that were about sexual content to ensure that the videos watched would be expected to have sexually suggestive content in them. The hashtags that were used were words with sexual connotations, along with their algospeak alternates, as found by Steen et al.. Algospeak and the traditional spelling of each word were chosen, so hashtags are expected to have sexually suggestive content since they are sexual in nature. Each of these words was searched as a hashtag on the TikTok app, and the three hashtags with the most views were chosen as they had the largest amount of exposure on the app.

Traditional Spelling	Views	Algospeak	Views
#sexworker	815.1K	#accountant	9.6B
		#sw	10.0B
#breasts		#b00bs	
		#🍒	1.6B
#porn		#corn	12.4B
		#🌽	589.7M
(no traditional spelling)		#fakebody	135.8B
#sex	8.5B	#shmex	965.5K
		#seggs	3.0B
		#🌶️	557.3M
#stripper	2.5B	#skripper	2.0B
		#Stripper	
#horny		#h0rny	
#femalegenitals	3.6M	#😸	803.7M
#malegenitals	1.6M	#🍆	
#pornstar		#🌽⭐	261.8M
#ejaculation		#💦	263.0M
		#🫧	1.5B
#pornhub		#🟠⚫	
#nipples		#nipnop	8.5M
#butt		#🍑	8.2B

Table 1: List of hashtags searched on TikTok's in its hashtag search section. Sections are divided by their traditional spelling, the algospeak alternatives of the word, and then the number of views under each hashtag with K standing for thousand, M standing for million, and B standing for billion. The hashtags highlighted in green are the chosen hashtags, and the ones in red were disqualified

The three hashtags originally chosen were #fakebody, which is used in videos with partial nudity to try and prevent it from being taken down, with 135.8 billion views, #corn,

algospeak for porn, with 12.4 billion views, and #SW, algospeak for sex worker, with 10 billion views (Picou 2022). After further inspection of the hashtag #corn, this hashtag was disqualified as most of the videos were about the grain corn and not its algospeak equivalent of porn and replaced with the next highest viewed hashtag, #accountant. This was also disqualified due to a majority of the videos under the hashtag being about managing money as opposed to the algospeak alternative, sex workers. The next viewed hashtag, #sex with 8.5 billion views, was used instead. A similar process was done with the hashtag #SW, as most videos were about the Star Wars franchise and not sex workers, being replaced by the hashtag #🍑 with 8.2 billion views. This led to the three analyzed hashtags being #fakebody, #sex and #🍑. Some hashtags were taken down by TikTok and unable to be searched, such as #breasts and #PornHub, so there is no view count associated with these hashtags.

Three TikTok accounts were utilized to test the effectiveness of the filters. All three accounts created ensured that no automatic age-based restrictions were put on the accounts by TikTok, their country set to the United States, and three email addresses from Google that were created for this study. Account A had no moderation settings turned on, Account B had Restricted Mode on, and Account C had Keyword Search on. The keywords that were inputted into Keyword Search are the same words that were first used to determine the hashtags being used, as featured in Table 1.

Each account watched the first 100 videos under each hashtag, with all videos being watched in full, and no user interaction, such as liking or following the account creator, to ensure that the TikTok algorithm recommending these videos is minimally influenced. To ensure an equal amount of videos coded visually and audibly, videos that did not feature any

kind of sound were skipped and not counted towards the 100 video count. These videos were also screen recorded using the screen recorder provided by Apple for rewatchability.

After collecting 100 videos for each of the three hashtags, they were coded visually and auditorily based on frequency. The visual codes were those Soriano-Ayala et al. used in their study about sexually suggestive content on TikTok: garments that emphasize the chest, garments that expose the neckline and/or the back, clothing with words/illustrations/images with erotic/sexual/aggressive and/or crude content, styles with underwear or swimwear visible under street clothes, seductive dancing (sensual movements), seductive gestures (lip biting, sticking out the tongue), very short clothing, and written sexual or degrading language (Soriano-Ayala et al., 2022). The auditory codes were those Wright and Rubin used in their study about sexual content in song lyrics: mentions of sexual behavior and body language (e.g. intimate touch, hand gestures to sexual acts), sexual language (e.g. talk about sexual encounters, advice regarding sex), and demeaning messages (e.g. objectification of women, sexual violence) (Wright & Rubin, 2016).

A second round of coding was done to ensure the validity of the codes, as only one researcher coded the videos. During the second round of coding, emergent codes were used to group videos into categories. After the data was quantitatively coded, basic descriptive statistics were used and a two-factor ANOVA statistical test to determine if there was significant difference in the number of codes shown based on video type, indicating the effectiveness of these filters.

This research fills a gap as there has been no other research focused on these opt-in content moderation settings on TikTok. The testing of these filters allows comparisons to be drawn regarding the amount of codes shown based on account type, and then statistical tests to

emphasize if there is a significant difference between the amount of codes shown based on account type. A significantly different amount of these frequency based codes suggest that these filters do influence the amount of sexually suggestive content shown in the videos, while a nonsignificant difference would suggest that these filters all show the same amount of sexually suggestive content.

Results

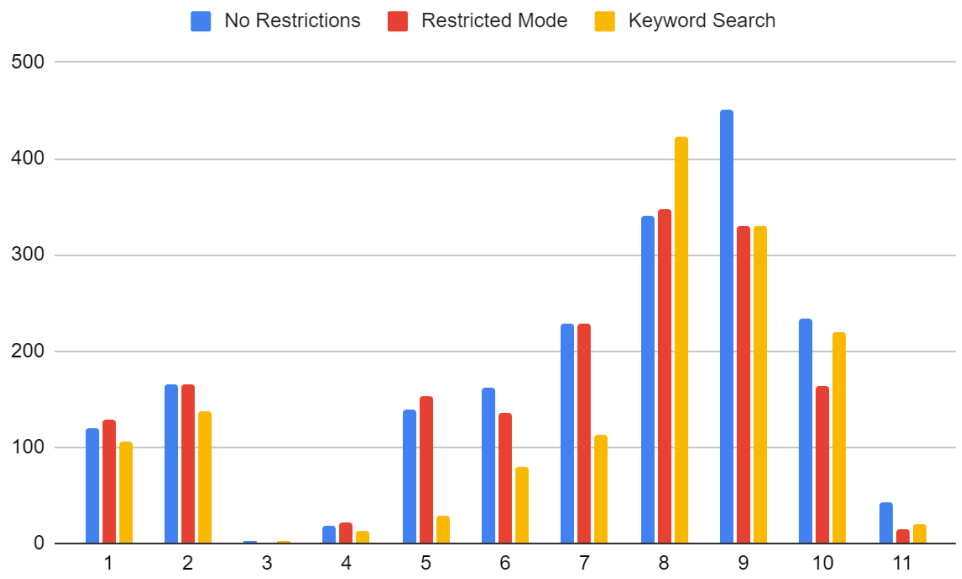
Quantitative Codes

In total, 900 TikToks were collected and coded, with a secondary round of coding done to ensure coding accuracy. Table 2a shows the quantitative results based on hashtag and user account settings

Restriction	**Hashtag**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
None	#fakebody	48	56	0	16	77	37	89	5	74	11	5
None	#sex	34	63	2	1	11	47	30	285	260	204	17
None	#🍑	37	47	1	2	51	78	103	50	116	16	10
Restricted Mode	#fakebody	59	78	0	18	106	21	118	4	31	1	1
Restricted Mode	#sex	48	50	0	2	1	31	25	293	189	145	10
Restricted Mode	#🍑	31	37	1	1	47	84	86	50	110	18	4
Keyword Search	#fakebody	30	42	0	9	10	15	27	3	54	16	1
Keyword Search	#sex	55	55	0	0	0	39	31	403	227	198	10
Keyword Search	#🍑	20	41	2	4	19	25	54	16	49	6	9

Table 2a: Table representing frequencies of codes found for each account (row) with different settings and hashtags viewed. Codes are (1) Garments that emphasize the chest. (2) Garments that expose the neckline and/or back. (3) Clothing with words, illustrations, and/or images with erotic/sexual/aggressive and/or crude content. (4) Styles with underwear or swimwear under street clothes. (5) Seductive dances (sensual movements). (6) Seductive gestures (lip biting, sticking out the tongue). (7) Very short clothing. (8) Written sexual or degrading language. (9) Mentions of sexual behavior/body language. (10) Sexual language. (11) Demeaning messages

In the account without restrictions, the videos under the #fakebody hashtag had a total of 418 codes, 328 visual and 90 auditory, #sex had a total of 960 codes, 479 visual and 481 auditory, #🍑 had a total of 513 codes, 369 visual and 144 auditory. The account with Restricted Mode turned on had 437 total codes under #fakebody, 404 visual and 33 auditory, #sex with a total of 794 codes, 450 visual and 344 auditory, and #🍑 with 469 codes, 337 visual and 132 auditory. The account with Keyword Search turned on had a total of 207 codes under the hashtag #fakebody, 136 visual and 71 auditory, #sex with a total of 1,018 codes, 583 visual and 435 auditory, and #🍑 with 245 total codes, 181 visual and 64 auditory.



Graph 2b: Table representing frequencies of codes found for each account. Codes are (1) Garments that emphasize the chest. (2) Garments that expose the neckline and/or back. (3) Clothing with words, illustrations, and/or images with erotic/sexual/aggressive and/or crude content. (4) Styles with underwear or swimwear under street clothes. (5) Seductive dances (sensual movements). (6) Seductive gestures (lip biting, sticking out the tongue). (7) Very short clothing. (8) Written sexual or degrading language. (9) Mentions of sexual behavior/body language. (10) Sexual language. (11) Demeaning messages

Looking at the amount of codes based on account type as shown in Graph 2b, there are no major skews in the frequency of codes for a majority of the visual code, with most of them having a similar amount of appearances based on account type. The only major difference is

between codes 4, 5, 6, and 7, where the Keyword Search account had noticeably less instances of these codes. These patterns were further shown below in Graph 2b, which shows the frequency of each code found in the videos based on the account settings used to collect videos.

In total the videos under the account without restrictions featured 1,891 codes, 1,176 visual and 715 auditory, the account with Restricted Mode had 1,700 codes, 1,191 visual and 509 auditory, and the account with Keyword Search had 1,470 total codes, 964 visual and 506 auditory. A two-factor ANOVA test without replication was performed to determine any significant differences between the amount of codes present based on the hashtag and account settings of the videos. The results of this test are shown below (Table 3).

Source of Variatio	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	2.962466667	2	1.481233333	0.8308310741	0.4991511579	6.94427191
Columns	59.315	2	29.6575	16.63503786	0.01151859298	6.94427191
Error	7.131333333	4	1.782833333			
Total	69.4088	8				

Table 3: Table showing results of a two-factor ANOVA test without replication using the data collected through the codes. The ‘Rows’ row represents the different content moderation settings, and the ‘Columns’ row represents the different hashtags

This statistical analysis accepts the null hypothesis that there is no significant difference in the amount of codes shown based on moderation settings, as the p-value is approximately 0.50 for ‘Rows’, and greater than 0.05. The second null hypothesis for this test, that there is no significant difference in the amount of codes present based on the different hashtags, is rejected by statistical tests as the p-value is 0.01 for ‘Columns’ and lower than 0.05.

Source of Variatio	SS	df	MS	F	P-value	F crit
Rows	4.282633333	2	2.141316667	1.347743079	0.4259409851	19
Columns	35.96601667	1	35.96601667	22.6369835	0.04144836827	18.51282051
Error	3.177633333	2	1.588816667			
Total	43.42628333	5				

Table 4: Table showing results of a two-factor ANOVA test without replication using the data collected through the codes. The 'Rows' row represents the different content moderation settings, and the 'Columns' row represents the different types of codes, auditory and visual.

A secondary ANOVA test without replication was performed to test another factor, this being the type of codes featured being visual or auditory as seen in Table 4. This statistical analysis accepts the null hypothesis that there is no significant difference in the amount of codes shown based on moderation settings yet again, as the p-value is approximately 0.43 for 'Rows'. The second null hypothesis for this test, that there is no significant difference in the amount of codes present based on the code type, either auditory or visual, is rejected by statistical tests as the p-value is 0.01 for 'Columns' and lower than 0.05.

Qualitative (Emergent) Codes

In addition to quantitative codes, emergent coding was used to categorize the types of videos shown. The codes that emerged for the videos were 'Dancing', videos featuring users dancing to music, 'Comedic (Sexual)', comedic videos with sexual themes, 'Comedic (Nonsexual)', comedic videos without sexual themes, 'Outfit', videos aimed at showing off the user's outfit choices or physique, 'Workout', videos consisting of fitness advice or workout routines, 'Sex Ed/Tips', videos consisting of educative content and advice regarding sexual activities, 'Podcast/Interview', videos consisting of user giving opinions relating to sexual themes, 'Miscellaneous Sexual', videos consisting of sexual themes but not fitting in any of the previously defined codes, and 'Unrelated', videos without sexual themes are unrelated to the

hashtag the video is under. The amount of videos under each category based on account type and hashtag is shown below (Table 5).

Restriction	**Hashtag**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
None	#fakebody	29	7	11	35	2	0	0	2	14
None	#sex	3	31	2	13	0	11	16	10	14
None	#🍑	13	32	7	10	19	2	3	2	2
Restricted Mode	#fakebody	23	8	32	16	1	0	0	1	19
Restricted Mode	#sex	1	30	6	2	0	19	5	15	19
Restricted Mode	#🍑	13	30	6	9	17	2	2	3	18
Keyword Search	#fakebody	6	7	66	11	0	0	0	1	9
Keyword Search	#sex	1	26	6	2	1	18	20	8	15
Keyword Search	#🍑	7	11	9	31	24	0	1	1	16

Table 5: Table representing amount of videos under each hashtag and moderation setting categorized under the categories (1) Dancing, (2) Comedic (Sexual), (3) Comedic (Nonsexual), (4) Outfit, (5) Workout, (6) Sex Ed/Tips, (7) Podcast/Interview, (8) Miscellaneous Sexual, (9) Unrelated

Under the account with no restrictions turned on, the most prevalent video types were ‘Comedic (Sexual)’ with 70 of the 300 videos fitting under the category, followed by ‘Outfit’ with 58 videos, and ‘Dancing’ with 45 videos, indicating that these were the most common types of videos under the account. The account with Restricted Mode turned on had ‘Comedic (Sexual)’ as the most prevalent video type with 68 instances, ‘Unrelated’ as the second with 56, and ‘Comedic (Nonsexual)’ with 44 instances as the third. The account with Keyword Search had ‘Comedic (Nonsexual)’ as the most frequent video type with 81 instances, followed by ‘Comedic (Sexual)’ and ‘Outfit’, both having 44 instances.

Discussion

This study was designed to determine the effectiveness of different content moderation settings on TikTok at moderating sexually suggestive content when they are placed in an environment that is likely to feature this content.

After examining the results from Tables 3 and 4, it can be concluded that there was no significant difference in the amounts of codes shown based on the account's moderation type. The p-values for the account type in the ANOVAs are all over 0.05, showing that there is no significant difference between the number of codes shown based on the moderation type. There has been a significant difference noted based on the hashtag the videos were under, #fakebody, #sex or #🍑, as seen in Table 3 with the hashtag type having a p-value of 0.01, and a significant difference based on the type of code, either visual or auditory, as shown in Table 4 with code type having a p-value of 0.04. These illustrate that the hashtag the videos were collected under and the type of codes, either auditory or visual, coded in each video caused a significant difference in the number of codes shown, but there was no significant difference based on moderation type. This leads to the conclusion that in this scenario, the two available opt-in moderation settings offered by TikTok were not effective at moderating sexually suggestive content when they were placed in an environment where this content is likely.

The visual trends in the videos can also be examined, namely the category the videos were sorted in as seen in Table 5. The video categories that focused on sexual themes were 'Comedic (Sexual)' in column 2, 'Sex Ed/Tips' in column 6, and 'Miscellaneous Sexual' in column 8. Totaling these numbers shows that the account without restrictions had 97 inherently sexual videos in total, the account with Restricted Mode had 108, and Keyword Search had 72. There is not a large variation in the amount of inherently sexual videos shown between the

account without restrictions and the one with Restricted Mode turned on, showing that there was not a large variance in the videos shown and both featuring an equal amount of codes. Keyword search has approximately 80% inherently nonsexual videos, offering that this mode may expose the user to less inherently sexual videos despite not resulting in a significant difference in the number of codes shown.

Filling the Gap

There has been no previous research done about testing the effectiveness of optional content moderation policies on social media as a whole, and none for TikTok. This paper aims to begin filling this gap, as it starts the process of examining the effectiveness of these content moderation settings on TikTok for one type of content, sexually suggestive content. The lack of a significant difference in the amount of codes shown based on account type indicates that these settings may not be entirely effective when placed in settings that have a high likelihood of having sexually suggestive content.

Despite the lack of a clear set of papers that can be connected to this study, it can add to general trends regarding sexually suggestive content observed by other researchers. These findings disprove Young's findings in 2021, where he proposed that the content moderation methods TikTok uses, a mix of humans and computer algorithms moderating content, would lead to difficulty distinguishing between sexually suggestive and sexually educative content, leading to sexually educative content being falsely labeled as sexually suggestive and wrongly taken down. As seen in Table 5, specifically column 6, when grouping the amount of sexually educative videos by account type, the account with no restrictions had 13 videos, the account with Restricted Mode turned on had 21 videos and the account with Keyword Search had 18 videos. There was no decrease in the number of videos with sexually educative content when

moderation settings were turned on, going against Young's prediction that the amount of these videos would decrease when moderation settings are turned on. This research also supports Steen et al.'s findings regarding algospeak, as some videos did feature algospeak in a written and spoken form when mentioning sexually suggestive or explicit elements. However, some videos did mention sexual elements in plain English or other languages, showing how users can also use plainly written words to post about sexual content.

Implications

With there being no significant difference in the amount of sexually suggestive codes shown based on the account moderation settings, the results indicate that there was no significant difference in the amount of sexually suggestive content shown to the user based on the moderation type. These results show that these filters were not effective at filtering sexually suggestive content when there is a high likelihood of being exposed to it for this research study. Adolescents using moderation settings would be shown the same amount of sexually suggestive content as those who do not use these settings, putting both groups equally at risk of potentially having more sexual thoughts and behaviors due to this exposure.

These findings could call into question the overall effectiveness of these moderation policies at filtering this type of content, as there was not a significant difference in the amount of codes shown based on account type. As TikTok markets these opt-in moderation features largely towards parents, it could cause parents to reconsider allowing their child on the app regardless of what settings they are using on their child's account..

Limitations

There was only one researcher conducting this project, so only one person coded the collected videos. While a second round of coding was done to try and minimize any personal

biases, there is still a possibility of bias impacting the frequencies of codes as there was no one to cross reference these codes with. Furthermore, the videos were collected over the span of a month, so the algorithm could have recommended different videos at the different times they were collected. Additionally, the moderation settings on TikTok are more tailored towards the 'For You Page', a content feed on TikTok that automatically recommends users videos, as opposed to the search bar as this study used. A more appropriate method of testing the effectiveness would have been through the 'For You Page', as opposed to certain hashtags.

Areas of Future Research

Future researchers can explore the effectiveness of these filters further. TikTok considers other elements to be mature, such as drug and alcohol use and violent content, and thus moderated under these opt-in moderation settings. Examining different types of mature content under the moderation of Restricted Mode and Keyword Search, would show if certain types of content are more effectively moderated than others. These moderation settings are also intended more for the 'For You Page' as opposed to looking for videos under hashtags. Improving the original methods, perhaps by interacting with content on the 'For You Page' to try and have mature content be recommended by the algorithm, and testing the moderation settings on the 'For You Page' will further illustrate the relative effectiveness of these filters in a way that closely resembles how they are used by users.

Works Cited:

Anderson, M., Faverio, M., Gottfried, J. (2023, December 11). Teens, social media and technology 2023. Pew Research Center: Internet, Science & Tech. https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/

Ayu, M., Izhar, T. A. T., Shahibi, M. S., & Kamaruzzaman, M. R. S. (2022). Hashtags: Social Media Community Information Dissemination Ultimate Tools. International Finance and Banking, 9(1), 18. https://doi.org/10.5296/ifb.v9i1.20016

Braams, B. R., Davidow, J. Y., & Somerville, L. H. (2018). Developmental patterns of change in the influence of safe and risky peer choices on risky decision‐making. Developmental Science, 22(1). https://doi.org/10.1111/desc.12717

Bucknell Bossen, C., Kottasz, R. (2020). Uses and gratifications sought by pre-adolescent and adolescent TikTok consumers. Welcome to the Kingston University Research Repository - Kingston University Research Repository. https://eprints.kingston.ac.uk/id/eprint/47303/

Calderón, D., Kuric, S., Sanmartín, A., & Megías, I. (2021). Barómetro Jóvenes y Tecnología. Trabajo, estudios y prácticas en la incertidumbre pandémica. Centro Reina Sofía sobre Adolescencia y Juventud, Fad, 2021, e507815. https://doi.org/10.5281/zenodo.507815

Christenson, P. G., de Haan-Rietdijk, S., Roberts, D. F., & ter Bogt, T. F. M. (2018). What has America been singing about? Trends in themes in the U.S. top-40 songs: 1960–2010. *Psychology of Music*, *47*(2), 194–212. https://doi.org/10.1177/0305735617748205

Coker, A. L., Richter, D. L., Valios, R. F., KcKeown, R. E., Garrison, C. Z., & Vincent, M. L. (1994, November). Correlates and Consequences of Early Initiation of Sexual

Intercourse. Wiley Online Library. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1746-1561.1994.tb06208.x

“Community guidelines enforcement report”. TikTok. (2023, December 13). https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2023-3/

Coyne, S., Ward, M., Kroff, S., Davis, E., Holmgren, H., Jensen, A., Erikson, S., & Essig, L. (2018). Contributions of Mainstream Sexual Media Exposure to Sexual Attitudes, Perceived Peer Norms, and Sexual Behavior: A Meta-Analysis. Journal of Adolescent Health, 64(4).

Evli, M., Albayrak, E., Şimsek, N., & Uzdil, N. (2023). Mindfulness, Psychological Well-being, Social Media Use, and Internet Use Time among Adolescents: A Structural Equation Modeling. Bağımlılık Dergisi, 24(4), 407–416. https://doi.org/10.51982/bagimli.1106080

Lewis, L., Somers, J. M., Guy, R., Watchirs-Smith, L., Skinner, S. R. (2018, June 21). “I see it everywhere”: Young Australians unintended exposure to sexual content online. CSIRO PUBLISHING. https://www.publish.csiro.au/sh/Fulltext/SH17132

Karsay, K., & Matthes, J. (2016). Sexually Objectifying Pop Music Videos, Young Women’s Self-Objectification, and Selective Exposure. Communication Research, 47(3), 009365021666143. https://doi.org/10.1177/0093650216661434

Kaynak, B. D. (2017). The Effects of Sexually Explicit Media Content on Adolescents and Some Intervention Methods. Nesne Psikoloji Dergisi, 5(10). https://doi.org/10.7816/nesne-05-10-04

Kharlamova, D., & Erofeeva, M. (2022). On The Issue Of Formation Of Behavior Of Adolescents On The Internet. The European Proceedings of Social & Behavioural Sciences, 128(2357-1330). https://doi.org/10.15405/epsbs.2022.11.49

Lebedíková, M., Mýlek, V., Subrahmanyam, K., & Šmahel, D. (2022). Exposure to Sexually Explicit Materials and Feelings after Exposure among Adolescents in Nine European Countries: The Role of Individual Factors and Social Characteristics. Archives of Sexual Behavior, 52. https://doi.org/10.1007/s10508-022-02401-9

"Our approach to content moderation". (2023, January 19). TikTok. https://www.tiktok.com/transparency/en-us/content-moderation/#:~:text=Automated%20moderation%20technology&text=These%20systems%20look%20at%20a

Paasonen, S., Jarrett, K., & Light, B. (2019). NSFW: Sex, Humor, and Risk in Social Media. In direct.mit.edu. The MIT Press. https://direct.mit.edu/books/monograph/4565/NSFWSex-Humor-and-Risk-in-Social-Media

Peters, J. (2020). Sexual Content and Social Media Moderation. Washburn Law Journal, 59, 469. https://heinonline.org/HOL/LandingPage?handle=hein.journals/wasbur59&div=30&id=&page=

Robb, M., & Mann, S. (2022). Teens and Pornography. In Pew Research Center. Pew Research Center. https://www.commonsensemedia.org/sites/default/files/research/report/2022-teens-and-pornography-final-web.pdf

"Sensitive and Mature Themes". (2023, March 8). TikTok. https://www.tiktok.com/community-guidelines/en/sensitive-mature-themes/

Setyawati, R., Hartini, N., & Suryanto, S. (2020). The Psychological Impacts of Internet Pornography Addiction on Adolescents. Humaniora, 11(3), 235–244. https://doi.org/10.21512/humaniora.v11i3.6682

Siswanto, S., Mar'ah, Z., Salsa Dila Sabir, A., Hidayat, T., Amir up Adhel, F., & Sitti Amin, W. (2022). Sentiment Analysis Using Naive Bayes with Lexicon-Based Feature on TikTok Application. Seminatic Scholar. https://www.semanticscholar.org/paper/Sentiment-Analysis-Using-Naive-Bayes-with-Feature-Siswanto-Mar'ah/1caf0c339fa88f4cda1350cf1f17bf027e6b467f

Soriano-Ayala, E., Bolo Díaz, M., Cala, V. C. (2022, July 13). Tiktok and child hypersexualization: Analysis of videos and narratives ... Taylor & Francis. https://www.tandfonline.com/doi/full/10.1080/15546128.2022.2096734

Steen, E., Yurechko, K., & Klug, D. (2023). You Can (Not) Say What You Want: Using Algospeak to Contest and Evade Algorithmic Content Moderation on TikTok. Social Media and Society. https://doi.org/10.1177/20563051231194586

Stoddard, R. E., Pelletier , A., Sundquist, E. N., Haas-Kogan, M. E., Kassamali, B., Huang, M., Johnson, N. R., & Bartz, D. (2023, October 4). Popular contraception videos on TikTok: An assessment of content and topics. Science Direct. https://sciencedirect.com/science/article/abs/pii/S0010782423004067?fr=RR-2&ref=pdf_download&rr=826cc1241cfdc402

Taylor, S. H., & Choi, M. (2022). An Initial Conceptualization of Algorithm Responsiveness: Comparing Perceptions of Algorithms Across Social Media Platforms. Social Media + Society, 8(4), 205630512211443. https://doi.org/10.1177/20563051221144322

Te'eni-Harari, T., Lehman-Wilzig, S., Lampert, S. I. (2007, September). Information Processing of Advertising among Young People: The Elaboration Likelihood Model as Applied to Youth. ResearchGate. https://www.researchgate.net/publication/240798185_Information_Processing_of_Advertising_among_Young_People_The_Elaboration_Likelihood_Model_as_Applied_to_Youth

TikTok. (2023). About | TikTok - Real Short Videos. Tiktok.com; TikTok. https://www.tiktok.com/about?lang=en

"User safety." (n.d.). TikTok; TikTok. https://support.tiktok.com/en/safety-hc/account-and-user-safety/user-safety

Wesche, R., Kreager, D. A., Lefkowitz, E. S., & Siennick, S. E. (2017, February 10). Early sexual initiation and mental health: A fleeting association or enduring change?. Journal of research on adolescence : the official journal of the Society for Research on Adolescence. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5546176/#:~:text=The%20transition%20to%20sexual%20intercourse,sexual%20initiation%20and%20internalizing%20symptoms

Wright , C., Qureshi, E. (2015, July 15). The Relationship Between Sexual Content in Music and Dating and Sexual Behaviors of Emerging Adults. ResearchGate.

https://www.researchgate.net/publication/273990137_The_Relationship_Between_Sexual_Content_in_Music_and_Dating_and_Sexual_Behaviors_of_Emerging_Adults

Wright, C. L., & Rubin, M. (2016). “Get lucky!” Sexual content in music lyrics, videos and social media and sexual cognitions and risk among emerging adults in the USA and Australia. Sex Education, 17(1), 41–56. https://doi.org/10.1080/14681811.2016.1242402

Wright, C., Toro Arenas, M., Martinez, P., McMullen, K., & Philip, R. (2018). “Love me!” Examining the effect of music censorship on sexual priming and sexual cognitions. Media Psychology Review. https://mprcenter.org/review/love-me-examining-the-effect-of-music-censorship-on-sexual-priming-and-sexual-cognitions/

Young, G. K. (2021). How much is too much: the difficulties of social media content moderation. Information & Communications Technology Law, 31(1), 1–16. https://doi.org/10.1080/13600834.2021.1905593

Merriam-Webster. (n.d.). Hashtag. In Merriam-Webster.com dictionary. Retrieved January 12, 2024, from https://www.merriam-webster.com/dictionary/hashtag

Picou, S. (2022, May 21). What does fake body mean on TikTok? The US Sun. https://www.the-sun.com/tech/5389435/what-does-fake-body-mean-on-tiktok/