AMENDED Exhibit 510

PLAINTIFFS' OMNIBUS OPPOSITION TO DEFENDANTS' MOTIONS FOR SUMMARY JUDGMENT

Postmortem – WSJ Algorithm Investigation

Privileged and Confidential

Issue Overview

The Wall Street Journal ran a series of bot TikTok accounts to investigate TikTok's AI and For You Feed, with a particular interest in our approach to minor safety.

> Over the course of several months, The Wall Street Journal set up more than 100 TikTok accounts that browsed the app with little human intervention, including 31 accounts registered as users between ages 13 and 15. The Journal also created software to guide the accounts and analyze their behavior.
>
> The Journal assigned each of its 31 minor accounts a date of birth and an IP address. Most were also programmed with various interests, which were revealed to TikTok **only through lingering on videos with related hashtags or images and through scrolling quickly past the others**. Most didn't search for content and instead simply watched videos that appeared in their feed.
>
> **About a dozen of the Journal's 31 minor accounts ended up being dominated by a particular theme**... Even when the Journal's accounts were programmed to express interest in multiple topics, TikTok sometimes zeroed in on single topics and served them hundreds of videos about one in close succession.

Commented [1]: From WSJ: " Most findings come from the process of the bot watching, rewatching, or not watching - carefully looked at watch time and amount of time spent on videos"

Commented [2]: [redacted] sevre your longer term research this week to ensure to pull length of each VID, UID's watch time per VID, and number of times the bot UID watched each VID

Commented [3]: Unfortunately the archived data for video play doesn't have watch duration, but the other data are available

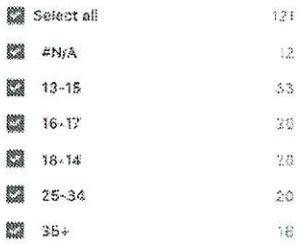
Rebecca Pober : [FINGERHEART] 2021-09-15 22:29:35

WSJ published two stories regarding their experiment:

- Published 7/21: https://www.wsj.com/video/series/inside-tiktoks-highly-secretive-algorithm/investigation-how-tiktok-algorithm-figures-out-your-deepest-desires
- Published 9/8: https://www.wsj.com/articles/tiktok-algorithm-sex-drugs-minors-11631052944

Videos viewed by WSJ Accounts - A Breakdown

After publication, we [HYPERLINK [redacted] that WSJ ran at least 6 devices and 121 accounts as part of their project. Breakdown of accounts registered age ranges:



Moderation Breakdown

[HYPERLINK [redacted]: Among all 120+ accounts ran by WSJ, all the content they viewed had a ~5.13% average violation rate and ~5.83% average NR rate. Current policy lacks guidance on nuanced areas for moderation (especially for Minor vs General).

[redacted]

Commented [4]: @Rebecca PoberWhat are the key takeaways from these numbers? It seems that the percentage of violating content is quite high.

Commented [5]: Added a tldr

Video Moderation Stats

For all of the videos seen by the WSJ accounts:

- Average number of times moderated - 3.3
- Highest number of times moderated - 49
- Highest number of policy titles chosen - 9
- Number of successful appeals - 1,162

External Perception Impact

[HYPERLINK [redacted]

Content Sent from WSJ

While working on the publications between June to August 2021, WSJ sent TikTok thousands of videos surfaced to two of their bot accounts, one surfaced a prominent amount of continuous drug-related content and another a prominent amount of continuous sexual and kink-related content. Prior to publication, these were the only two accounts for which we had data.

Commented [6]: Although the article also mentions the serving the bot accounts content promoting paid sex sites (likely OnlyFans), I don't see a lot of data indicating this. Did WSJ sent us anymore examples of this or is ANSA primarily focused on Kinks/Fetishes/Language?

Commented [7]: [redacted] tagged you in the spreadsheet of WSJ's examples. While WSJ did not send us many examples like this before publication, the published article does key in on ANSA-related content beyond just kinks/fetishes/language: "Several of the Journal's accounts were programmed to rewatch videos with sexual text or images, and soon fell into rabbit holes of sexual content." There's ANSA under hashtags like these (having Ops sweep): #knktok #spicycontentcreator #accountantlink #spicycontent Before the article published as a prevention measure for discoverability, we removed from Search auto-sug 18+ related keywords: <at type="1" href="https://byt[HYPERLINK [redacted] ' \h]ut the presence of OnlyFans content: [redacted] Unfamiliar with how we are thinking about the Search experience for <18 accounts, but Onlyfans terms also pop up in Search auto-suggest:
[image]
[image]
[image]

Account 1 – drug content

UID 6937396511799411718

The WSJ programmed "set interests" for one bot around topics: kink, bodybuilding, fitness, weight loss, marijuana, sex and 18+ content, stripper content, modeling, LGBTQ, and cosplay.

On July 20, WSJ sent TikTok 3,000 videos surfaced to this bot account that was almost exclusively served drug-related content beginning around video 2,271.

Moderation

The large majority of the videos surfaced to this bot account did not violate our policies, but 60% of the drug-related videos should have been Not Recommended (NR). The root cause was that the global TnS team's "mention of drugs" policy was worded differently in the US market and led to many of videos being approved instead of NR, specifically around the glorification or mention of cannabis. This was corrected and the policy was rewritten and cascaded to US moderation sites May 17, 2021. The content found by the WSJ was published between June 17, 2020 to March 11, 2021 before this policy iteration.

Commented [8]: So we won't clean past videos when we upgrade policy ? (just want to make sure I understand the current practice)

Commented [9]: Hi [redacted], at present no. We are doing a sweep for legacy content that we will NR but it is not something that we do after a policy iteration.

Prominence of Drug-Related Content

Word cloud of most common keywords used in Caption, Sticker Text, and OCR across the 3,000 videos surfaced to this bot account:

Account 2 – kink content

UID 6937430093322322949

On Aug 20, WSJ sent TikTok 1,278 videos surfaced to another one of their bot accounts that was almost exclusively served content focused on sex and kinks after a few hundred videos.

Quote of this account from WSJ article:

> "After signing up for an account on TikTok, a 13-year-old user started by searching the app for "onlyfans"—the name of a site known for hosting adult entertainment—then watched a handful of videos in the results, including two selling pornography."
> - https://www.wsj.com/articles/tiktok-algorithm-sex-drugs-minors-11631052044

Moderation

441 (35%) of the videos violated our policies, mostly for sexually explicit language. 253 of these violative videos were missed by moderation:

[redacted]

Policy interpretation might be a primary cause of the mismoderation.

Background

The videos sent by WSJ were published between June 18, 2020 through March 11, 2021.

- There were at least six US/CA QA alignments conducted for R1 and R2 between June 18, 2020 through March 11, 2021 across sexually explicit language, allusion to minor sexual activity, normalization of pedophilia, digital character content, and mention of drugs and other controlled substances.
- US/CA relevant policy cascading:
  - June 18, 2021: cascaded a policy reminder for "sexual explicit language" through policy refresher training
    - [HYPERLINK "[redacted]"]
  - August 13, 2021: re-incorporated guidance on "allusion to sexual activity" into policy refresher training

Prominence of Sex-Related Content

While only 35% of the overall content surfaced to this bot was violative or NR'd, 982 (77%) out of the 1,278 videos were related to sexual topics.

[redacted]

Word cloud of most common keywords used in Caption, Sticker Text, and OCR across the 1,278 videos surfaced to this bot account:

Timeline of Events

Date	Team	Event
5/27/2021	Global PR	initial outreach from WSJ

Commented [10]: For next steps, it would be worth taking a look at how this insight compares to the Badness comparison we did recently with GB ([HYPERLINK "[redacted]". TL;DR, we found that US typically overmoderates on sexually explicit language compared to GB. My hunch from the Badness analysis is that much of the content we "overmoderate" (violate) on is likely OK for >18 users, but not so much for <18. Would be good to see if that hunch is correct and see how that aligns with policy freshers and changes that <at type="0" hre@Tara Wadhwa" hr[redacted]

Commented [11]: @Amy Classenincreasing vis on this point. Even though numerous videos were false positives, that doesn't indicate it was not adult or sex related If you filter here: <at t[HYPERLINK "[redacted]" only False Positives, you can see in the "Why violative or not violative" column the type of content depicted – includes quite a bit

Commented [12]: [redacted] this is an area I'd like you to both explore in terms of 1. Whether

Commented [13]: I see from the video review that all 'mentions of fetishism' and educational content on

Commented [14]: @Amy ClassenI do think this is an issue of filtering and curation, not policy gaps. Based on some

Commented [15]: We should definitely consider filtering fetish related content out for our U16 users [redacted]

Commented [16]: Ok so it sounds like EDSA and 'mentioning fetish' content that is currently marked as

Commented [17]: Noting here that as we've thought about content classification we've focused a lot on using

Commented [18]: I met with [redacted] about Comment Policy, which is tangential, but I feel it is similar

Commented [19]: That is great feedback and something we should definitely consider as a newer-version of the 18

Commented [20]: In conversation with Wenjia last night, a keyword approach via Protected Mode was discussed.

Commented [21]: [redacted] egarding providing creators with an option, they already (informally) use

Commented [22]: Also, just wanted to highlight a point that@Julie de Bailliencourtmade earlier. A dedicated 18+

Commented [23]: I talked with [redacted] on this and we're aligned we dont' want an '18+' button yet until we actuall

Commented [24]: [redacted] agree fully, and I have been of the opinion that it's even more than a

Commented [25]: @Rebecca PoberHow about the timeline after the second report?

Commented [26]: [redacted]

5/28/2021	Global PR	first call with WSJ to hear what they were working on; PR handling doc created: [HYPERLINK [redacted]
6/4/2021	Global PR, Exec Leadership, TnS, GSO	call with WSJ to hear their findings; PR handling doc updated: [HYPERLINK [redacted] Lark group created with Vanessa, [redacted] Cormac, Wenjia, [redacted] PR informed [redacted] Wenjia, and [redacted] of the bot behavior
6/7/2021	Global PR	WSJ emailed more findings about the bot accounts: [HYPERLINK [redacted]
6/8/2021	Global PR	Cormac added [redacted] and Suzy to Lark group PR notified Shou of the two working WSJ pieces
6/9/2021	Global PR, TnS, PM	• WSJ given TAC tour with Michael Beckerman, Tara, [redacted] • US Safety advocates to Vanessa for more attention on harmful echo chambers that may form in response to the WSJ [HYPERLINK [redacted]
6/10/2021	TnS	TnS [HYPERLINK [redacted]]WSJ's claims that TikTok hosts accounts seeking to redirect traffic to sex sites using the hashtag #accountant and serves many of these (teenage) accounts content promoting paid sex sites, services and sex shops.
6/11/2021	PR	PR [HYPERLINK [redacted]]with WSJ reporter
6/21/2	TnS	WSJ sends 74 videos surfaced to their various bot account.

Commented [27]: @Rebecca PoberMichael Beckerman or [redacted]

021		Content surrounds sexual kinks and fetishes, drugs, drinking alcohol, and eating disorders. [HYPERLINK [redacted]
6/24/2021	PR	PR [HYPERLINK [redacted]] with WSJ reporter
6/30/2021	TnS	WSJ sends 1,000 videos from two "age 13" bot accounts. US Safety investigates • [HYPERLINK [redacted] served to one bot account • [HYPERLINK [redacted] served to other account [HYPERLINK [redacted]
6/30/2021	PR	PR responds to WSJ with findings from review of the 1,000 videos
7/19/2021	TnS	WSJ provides all[HYPERLINK [redacted] shown to the bot account that was prominently surfaced into **drug** content
7/21/2021	All	WSJ publishes video piece, primarily focused on a bot account that was prominently surfaced **depression** content https://www.wsj.com/video/series/inside-tiktoks-highly-secretive-algorithm/investigation-how-tiktok-algorithm-figures-out-your-deepest-desires/ ◦ [HYPERLINK

		[redacted] [HYPERLINK [redacted]]of bot account featured in video
8/20/2021	All	WSJ provides all [HYPERLINK [redacted] shown to the bot account that was prominently surfaced **sexual/fetish** content
8/25/2021	TnS	Review of the sexual/fetish bot account complete: [HYPERLINK [redacted] k
9/8/2021	WSJ	WSJ publishes written piece – particularly pointing out accounts dedicated to sexually oriented content, including violent sex acts https://www.wsj.com/articles/tiktok-algorithm-sex-drugs-minors-11631052944

Business Impact

Why do we need to solve this problem from a T&S perspective?

The safety of our user community is at the core of TikTok's mission to inspire creativity and bring joy. Gaps in our product protections for minors can particularly have downstream reputational and revenue impact. For example, one Wall Street Journal article has put approximately $20-30M revenue over the next 12 months at risk, in addition to losing approval to run ads in 20 markets (specific to [HYPERLINK [redacted]. Our business partners hold us to high standards on user safety and issues uncovered by the WSJ's investigation pose significant risk to the TikTok brand, impacting user growth and the ability to scale business partnerships.

Protecting the safety of minors is at the heart of all company goals and objectives. Minor users and their guardians should feel confident in TikTok's ability to safeguard them from harm and inappropriate content. Improving the minor experience will help to establish TikTok as a committed partner to drive positive change and safety throughout the platform.

Next Steps

1. Dedicated PM to ensure accountability and feedback loops across all efforts to ensure coordinated company effort

a. Collect all current proposed or live projects related to addressing the key issues highlighted by the WSJ escalation into a single project tracker to reduce redundancy and ensure collaboration and alignment when needed

Commented [28]: @Eric Han@Rebecca PoberIs this been done actually?

Commented [29]: @Rebecca PoberJust for live projects?

i. Review all projects to identify dependent work-streams or processes that are necessary to each projects success. Ex. If an algorithmic solution requires accurate moderation and labeling, then moderation accuracy will need to be reviewed to ensure that accuracy rates are high enough for the algorithmic solution to work

ii. Analyze identified dependencies to ensure functionality

2. Comments on Current Known Projects

- Sources: [HYPERLINK [redacted] [HYPERLINK

Current Team	Current Known Project	Proposed Projects and Comments
Search	[HYPERLINK [redacted]	[HYPERLINK [redacted]
Content Classification	[HYPERLINK [redacted]	Consider grey area content areas. Policy is working to build out a list of grey area/on the margins content that should be labeled L1, but ensure reliance on product interventions rather than policy.
Algo	Short-term: [HYPERLINK [redacted] [HYPERLINK [redacted]	Dedicated resources on filter bubbles, as filter bubble projects were deprioritized in 2020: [HYPERLINK [redacted]

[HYPERLINK [redacted]

Long-term:

[HYPERLINK [redacted]

[HYPERLINK [redacted]

Strategic Recommendations

Age-Appropriate Content Filtering and Age Appropriate Design

Starting with a simple question around what an age appropriate design looks like can help guide our resolution efforts. Our hypothesis is that we would like to develop a framework that is capable of recognizing minor accounts and is able to throttle back and restrict access to inappropriate content. However, there are considerations to be made around what we would consider inappropriate content for minors. This requires significant focus on search and discoverability, content classification efforts, filters and bubbles, and policy enhancements.

The Basics

1. **Short-Term**: Define what the child/ minor experience is supposed to be and identify where we need to build/ fix accordingly

2. **Medium-Long Term**: Determine what minors should and should not see. Develop and revise content classification system to attach a label in Round 1 that would prohibit inappropriate content from entering a minor's FYF

- Confirm how we identify minor accounts and dedicate resources to building age detection solutions
- Age Detection Features
  - Current Project Status (ON TRACK): We are rolling out our Underage Appeals test in the EU (aiming for Oct 14th launch) to test for 3 months. During which, we will also launch more aggressive detection strategies for minors. After the trial, we will expand globally.
- Launch Kids Mode globally

- Current Project Status (OFF TRACK): Targeting to test Kids Mode outside the US in in 2021, but due to the complex factors of content sourcing, policy, and VPC, we likely will not revisit launching until 2022.

- Restricted Mode Launch and Family Pairing Efforts
  - Current Project Status (UNKNOWN): [waiting for update from Lufan]

Commented [30]: Content classification beta launch and Family Pairing Efforts (see details in[HYPERLINK [redacted] current project status: we are planning to filter out content with L1/2/3 maturity tags for logged-out/Kids Mode and Restricted Mode users. And we'll upgrade Restricted Mode to content classification user settings and parents can adjust this settings for the linked teen account via Family Pairing.

Keyword-Based Dispersion

Based on an analysis of [HYPERLINK [redacted] we recommend starting to disperse content with **2 or more** keywords for L1/L2 users, which is about 50k videos per month. To address borderline content concentration in L1/L2 users' feed, we will try two options:

3. **Short-Term**: Apply protected mode on borderline content keywords and disperse for L1/ L2 users
   - Investigate 'Kink' community to better advise on the associated risk - Hunt Team
   - Work with P&P to implement dispersion logic for L1/L2 users and create SWL to facilitate expansion of keywords
   - Send cases for labeling to build the model training dataset
4. **Medium-Term**: Build a detection model to flag grey-area content and disperse for L1/L2 users
5. **Long-Term**: Content Classification (refer to Content Classification sub-section)

Search and Discoverability

Minor Safety Gaps

TikTok's search function allows users to locate content on virtually any topic. Current [HYPERLINK [redacted] are focused on *"wanting to raise awareness of potential uncomfortable search results and allow users to consume if they decide to"*. While this customizes the search and discoverability for adult users, it abandons the minor search experience. This can often be considered patient zero for a minor's exposure to inappropriate content and has downstream impact on bubbles and the volume of inappropriate content available to minor users.

6. **Short-Term**: Dedicate resources to minor search optimization efforts
7. **Medium-Long Term**: Align with age detection and content classification efforts to begin restricting search options for minors

Media Response Tooling and Resource Gaps

TikTok's tools and processes for limiting Search and Discoverability of high-risk trends in rapid response crises such as the WSJ article currently cause severe limitations to quickly

act upon violative content being promoted in Search to prevent negative headlines and keep the TikTok community safe, often impacting the brand and business. Currently, the process is highly manual, requiring multiple XFN teams with sometimes 12+ hours delay to appropriately . Search is also one of the most common tools journalists utilize to identify harmful content on TikTok and publish negative press.

8. **Short-Term:** [HYPERLINK [redacted]

9. **Medium-Long Term**: Ensure adequate Search and Discoverability resources in US timezone; invest in building long-term tools to control the Search and Discoverability pages.

Filter Bubbles

Currently, we don't have a way to identify when a user has entered into a bubble. Furthermore, we don't have a way to identify harmful bubbles. To effectively restrict a user's involvement and exposure to inappropriate content, we need to find a way to identify and break harmful bubbles and clusters. This also requires the consideration of what content is ok for adults, but not minors. What risks do we inherit when grey area bubbles surface across various verticals? When and if we do break harmful bubbles, do we negatively impact the adult user experience? For adult users, how do we balance and accurately measure benefits of bubbles such as product stickiness with potentially harmful safety risks?

10. **Short-Term:** Revitalize [HYPERLINK [redacted] around (1) detecting harmful filter bubbles (2) diversifying harmful bubbles when detected and (3) recall potentially problematic content to downrank or enqueue for moderation review

11. **Medium-Long Term**:

- **Identify cross-functional team to define and identify harmful filter bubbles**: Team should include Product/Algo, RD, T&S, and Security. In collaboration with peripheral stakeholders (e.g., Public Policy, Comms), T&S can help define what a harmful filter bubble is (e.g., eating disorders, pornography, self harm, misinformation) and build a spectrum of acceptable content and violative content that may exist within a bubble -- for example, an eating disorder bubble may include exercise videos (acceptable) and "proana" videos (violative).

- **Analysis on identified bubbles**: Once defined and identified, conduct analysis focused on areas such as (1) the number of users that are in bubbles (2) the engagement of those users (3) the make up of those users -- i.e., L5 rural male (4) the time it takes for a user to enter a bubble (5) the percentage of users that eventually leave a bubble (6) the percentage of violative versus accepted content within a bubble, etc.

- **Develop bubble interventions**: Once analysis is complete, develop interventions that can "bust" harmful bubbles when we identify that a harmful one may be forming. It has been suggested that we increase diversification efforts by either (1) uploading more

"harmless" content -- i.e., puppies, food and (2) resetting their feed back to an experience like the New User Feed.

- **Use our harmful bubble detection for moderator recall**: The formation of harmful bubbles on platform is a negative user experience and hurts our brand safety, though the benefit may be using this technology to help recall or detect this type of content before they amplify. For example, if we are able to harness the formation of harmful bubbles to (1) downrank content and (2) enqueue videos that surpass a certain model threshold to a moderator for review, we'll have a powerful detection tool that will be complementary to our computer-vision risk models and user reports

Content Classification

Content Classification is critical to elevate TikTok to the next level for minor protection. Providing an appropriate experience for minors is required by legal and regulatory.

12. **Short-Term**: Grey area content poses a significant risk to minors. Policy is working to build out a list of grey area content poses significant risk. Policy is working to build out a list of grey area/ "on the margins content" that should be labeled L1

13. **Medium-Term**: Ensure reliance on product interventions rather than policy

Model Improvements

14. **Medium-Long Term**: Improve the Ad Adjacency/ Brand Safety Verification model by supplying WSJ data set to train the model to avoid ad placement/ adjacency to bad content. This will reduce the need for third party verification (i.e. OpenSlate, IAS) and alerting of ads being run against adjacent to bad content for our business partners

- Understanding how good content is nested with bad bubbles - find breakdowns in logic, queues, vetting, and moderation that allows for badness to get on the FYF (Ads Policy)

Target Improvement Timelines

Keyword-Based Dispersion (November 2021)

Video-based Maturity Model v1 (EOY 2021)

Age-Appropriate Search Experience (EOY 2021)

Video-based Maturity Model v2 (EOQ1 2022)

Questions

[redacted]

Appendix

Use	Description	Document
Escalation handling	Full list of videos sent from WSJ	[HYPERLINK [redacted]
	Comms handling documents	[HYPERLINK [redacted]
		[HYPERLINK [redacted]
	Comms final statements	[HYPERLINK [redacted]
		[HYPERLINK [redacted]
	Sweeps of "accountant"-terms for paid sexual solicitation content	[HYPERLINK [redacted]
	Review of related hashtags to kink/fetish communities	[HYPERLINK [redacted]
	IARG PnP review of WSJ's drug bot account and next steps	[HYPERLINK [redacted]
		[HYPERLINK [redacted]
	IARG keyword	[HYPERLINK

	enforcement for drugs	[redacted] Audit of antidirt inconsistencies for drugs: [HYPERLINK [redacted]
Projects in Response to WSJ article		[HYPERLINK [redacted] [HYPERLINK [redacted]

Parking Lot

- **Recommendations**
  a. Content Diversity
    i. Sensitive content: Diversify content by instituting enforcement actions associated with dispersion in sensitive policy titles
    ii. Content shown to minors: Align Product and Policy teams around sensitive policy titles (e.g., mention of fetishism and drugs) to classify this content and filter from the For You Feeds of minors.
  b. The In-App Search Experience
    i. Evaluate the Search strategy from the TnS perspective – what should be searchable at all, what type of content should surface high on Discover page, should there be a different Search experience for minors vs. general users, etc.
    ii. ~~Evaluate how Search vv impacts what is surfaced to the FYP~~
    iii. ~~Align with keyword experts to ensure consistent experiences across keyword variants (For example, if "18+" is blocked from search, so should "eighteenplus", etc.)~~
  c. Policy Development & Risks
    i. Evaluate the primary policy categories and playbook rules central to the WSJ videos:

	Mismoderation Issues based on[HYPERLINK [redacted] **from WSJ**	**Suggested AI**
IARG	310 videos needed to be Not Recommended for **"mention of drugs"**	These videos were posted in March and a [HYPERLINK [redacted] was cascaded to moderation sites on 5/19/2021. Continue to monitor.
	26 videos needed to be Violated for **"depiction of drugs" (primarily cannabis)**	Align US and Global position and approach to cannabis policy development.
ANSA	206 videos needed to be Violated for **"sexually explicit language"**	Check calibration and training for "sexually explicit language" policy playbook rule
	At least 28 videos needed to be Not Recommended for **"mention of fetish"**	Clarify the "Mentioning a sexual fetish" exception to "Fetishism involving adult" policy playbook rule
MS	Fetishizing minors (e.g., "littlespace")	Minor Safety and ANSA teams should evaluate and clean up content where needed surrounding the ddlg, subdom, littlespace, and ageregression communities on platform – [HYPERLINK [redacted]

Issue Policy [redacted] All GPIOs

Issue PnP [redacted] All GIOs

Algo & Product [redacted]

- [HYPERLINK [redacted]

Detection

Aggressive Machine Learning detection (20% reduction in violations for minors) rolling out now, and planning to try more aggressive approaches this coming bimonth. This approach uses two techniques: first, if the Gandalf machine learning model believes that content violates our existing policies, we will not put it in kids' FYF. (This is similar to automoderation we already do for adults, but we will use more aggressive thresholds for children.). Had this strategy been in place, it would have eliminated some of the content that was missed by moderation. The other strategy we use is "dispersion" - if we believe this content is likely violative, we will put less of it in kids' FYF. (So, if we're pretty sure - we take it out; if we're less sure, we show less of it.)

In the coming bimonth, we will experiment with two variations on this approach. For one variation, we will simply use the same technique, but with even more aggressive thresholds (take out and disperse more content; need to determine impact on user engagement.) Second, we will try to specifically target some of the worse policies (Suicide, Self Harm, Eating Disorders, Dangerous Challenges, Drugs). If we think that content even might possibly violate these policies, we will remove or disperse it from the FYF especially aggressively for minors.

Content Classification

[HYPERLINK [redacted] \h]

- [redacted] Currently mapping 4.0 policy titles to maturity ratings
- This project is in progress now (still determining timeline - [redacted] can give details). This project will mark certain content as inappropriate for minors. The previous machine learning approach attempts to more aggressively enforce our existing policies; Content Classification looks specifically for content that is OK for adults or older teens, but not OK for younger users.
  - For example, analyzing existing [HYPERLINK [redacted] and assessing potential rating. The current Drug Model can be conservatively rated M3 (age 18+).

Commented [31]: This is a significant step in the right direction. We have to prioritize it to the top and have effective x-team collaboration. Good news is that a lot of ground work is completed.

Profanity Filtering

- Profanity filtering - partially rolled out. This might have removed some of the worst content, especially some content with kink, sexual or other explicit content.
  - [HYPERLINK [redacted] have been assigned

Mature Hashtag Filtering

"Mature" hash tag filtering - will trial filtering content with certain hashtags indicating adults only. If successful, we will consider making this official. The WSJ article mentioned specifically that some adult users are unofficially using hash tags to mark their content as inappropriate for minors, and we'll try filtering that content from minor users. Depending on what we learn, we will trial more official approaches, e.g. giving creators a button to mark their content as inappropriate, or suggesting that they add a certain hash tag.

- [redacted] @Amy Classen are working to filter the 'M3 - MATURE (18+ years)' from minors and aim to have a test complete on ANSA and Substances this bimonth once we create the models

Search

[HYPERLINK [redacted] [HYPERLINK [redacted]] are working together with the new investigations team to audit the keyword strategy and make recommendations for search. But, we're starting with inappropriate Minor safety + ANSA keyword combinations (ex: hot + 14 year old) but this is an area that can and should be reviewed by all issue verticals. It is a cross-issue problem and we have a lot of basic misspellings that are still missed. We need to start to move away from only using keywords.

Algo & Product

- [HYPERLINK [redacted]
- [HYPERLINK [redacted]

P0	Project 1: A solution to filter [redacted]	Label depressive content by TnS policy and disable depressive contents in FYP	Redo WSJ experiment and evaluate the filter to make [redacted]	[redacted] Ongoing

Commented [32]: I think this is a specific aspect of our Content Classification approach, and we should merge with Content Classification section above?

Commented [33]: Agree. This would be a subtask within the broader Content Classification initiative.

Commented [34]: Search is a topic Wenjia and I discussed at length. Here is a quick summary that was put together reflecting on our current gaps with regards to Search:[HYPERLINK [redacted]]forward, can we make sure to have folks on the US side involved as well -- including myself and <at type=[redacted]

Rebecca Pober : [FINGERHEART] 2021-09-15 02:11:58

Commented [35]: From a scalability perspective, this is a textbook definition of the need for a text-classifier. Have we explored the handling of queries with a text-classifier? For precedence: This is what SafeSerach mode on Google Search does. It checks for any violative searches: For ex: "12 yr old porn"; is illegal to search for in most countries in the world. SafeSearch classifies this string as violative and serves a hotline/relevant ngo contact info. Or searching for "how to commit suicide" gives a result of suicide prevention lines based on the region the user is querying from.

Commented [36]: We do have these classifies for ssh and for other violations we just don't serve the content. But MS is working with [redacted] on new messaging this bi-month BUT the problem is the keywords aren't comprehensive enough

Commented [37]: if it's a question of having enough keywords to train the classifier for MS, [redacted] built and maintains a keyword list for MS and we could work with them go get that list and train our classifier.

Commented [38]: It is now part of the safer toolkit from [redacted]. https://safer.[redacted]o/resources/csam-keyword-hub/ Let me know if this can unblock the model training for MS. I can get us in touch with the right folks a[redacted]

DOCUMENT METADATA

Bates Number: TIKTOK3047MDL-060-01110548-TIKTOK3047MDL-060-01110564

BEGATTACH: TIKTOK3047MDL-060-01110528

ENDATTACH: TIKTOK3047MDL-060-01110565

Attachment Count: 0

PRODVOL: TIKTOK3047MDL-060

Custodian: ERIC HAN

File Path: DOCS.ZIP/FILES/

Confidentiality Designation: CONFIDENTIAL

HASHVALUE: 07A1AB69522C74A20212B09DB6B78AB4

DocType: E-DOC ATTACHMENT

Author: REBECCA POBER

Create Date: 7/2/2021 12:00 AM

Last Modified Date: 12/22/2021 12:00 AM

LAST MODIFIED BY:

TRACK CHANGES:

COMMENTS:

HASHIDDENDATA:

Filename: POSTMORTEM Â€“Â WSJ ALGORITHM INVESTIGATION-DOCCNT45KQ6DAXFDZ1LEAEONKKC.DOCX

Title: POSTMORTEM Â€“Â WSJ ALGORITHM INVESTIGATION

DOCEXT: DOCX

Email From:

Email To:

Email CC:

Email BCC:

Received Date:

Sent Date:

TIMEZONE:

Subject:

THREADID:

REDACTIONS: N

REDACTION TYPE: PRIVILEGE