

# THE WALL STREET JOURNAL.

For access to articles, use login: wsjcontests@gmail.com; password: judges

---

## 2021 Philip Meyer Journalism Award, Questionnaire

### How TikTok Figures You Out

**1. Title of story(ies) or series and names of people, including and identifying freelancers, who worked on this story.**

“How TikTok’s Algorithm Figures Out Your Deepest Desires” and “How TikTok Serves Up Sex and Drug Videos to Minors”

The personnel involved were Rob Barry, Yoan Cart, Dave Cole, Jason French, Robert Libetti, Maureen Linke, William Mata, Frank Matt, Darnell Stalworth, Joanna Stern, Christopher S. Stewart, Kenny Wassus, Georgia Wells, and John West

**a. For a partnership or collaboration, please name each entity that took part in the investigation.**

N/A

**2. Date(s) published, aired, or posted.**

“How TikTok’s Algorithm Figures Out Your Deepest Desires” ran on July 21, 2021 and “How TikTok Serves Up Sex and Drug Videos to Minors” ran on Sept. 8, 2021.

**3. Topic and synopsis of story or series, including major findings.**

“How TikTok’s Algorithm Figures Out Your Deepest Desires”: A Wall Street Journal investigation found that TikTok only needs one important piece of information to figure out what you want: the amount of time you linger over a piece of content. Every second you hesitate or rewatch, the app is tracking you.

“How TikTok Serves Up Sex and Drug Videos to Minors”: TikTok can quickly drive minors—among the biggest users of the app—into endless spools of content about sex and drugs.

**4. How the story got started (tip, assignment, etc.)?**

Assignment.

**5. Was your work in any way based on another news organization's previously published or aired report? If yes, please provide the name and publication or air date.**

No.

**6. Results/impact (if any).**

Sen. Amy Klobuchar cited Wall Street Journal reporting in a recent Senate Commerce Committee panel that she said showed “TikTok’s algorithm can push young users into content glorifying eating disorders, drugs, violence.”

“Have you stopped that?” Ms. Klobuchar asked a representative for the company.

[Sen. John Thune said](#) the Journal's reporting helped make the case for a bipartisan bill that would, among other things, require social media platforms to give users an alternative to opaque algorithms that curate the content they see.

<https://www.thune.senate.gov/public/index.cfm/press-releases?ID=0CA78D6E-C0A8-4BDB-9AA9-1900238810E5>

“Congress needs to be heard from in this space, particularly with respect to the use of algorithms and the way that users are manipulated,” he said.

The Journal also shared with TikTok a sample of 974 videos about drugs, pornography and other adult content that were served to minor accounts the Journal created for the story—including hundreds shown to single accounts in quick succession.

Of those, 169 were removed from the platform before the Journal shared them—whether by their creators or TikTok couldn’t be determined. Another 255 were removed after the Journal shared them with the company, among them more than a dozen portraying adults as “caregivers” entering relationships with people pretending to be children, called “littles.”

Following the Journal’s reporting, as an effort to protect minors, TikTok said that it would begin adding “distressing content” warning labels to harmful content.

**7. The following questions deal with the social science research method(s) you used; answer any that apply to your work:**

**a. Describe in detail the social science research method(s) used.**

Though American teenagers average many hours a day on social media, the platforms are largely unregulated, and their algorithms opaque.

Nowhere is that more true than with TikTok, the world's fastest-growing social media application. The platform serves an endless stream of personally customized videos, which are uploaded at a rate of many thousands a minute. The user experience is almost entirely algorithmically driven—and the platform has not been transparent about how it works.

To better understand what TikTok was serving its users, Journal reporters built an infrastructure for creating automated accounts that interacted with the platform and recording every video they were served.

In all, the Journal deployed more than 100 automated accounts, or bots, and collected hundreds of thousands of videos.

The bots were programmed to linger on different types of videos based on a broad array of signals, thus sending TikTok a signal of their interest.

Many bots used keywords or hashtags in the video description and author bios to determine how long to watch each video. Others passed thumbnails from the videos through machine learning image classifiers, and dwelled on videos if they matched certain classifications.

In both cases, reporters found that TikTok's algorithm would quickly serve a great deal more of the kinds of content the bots paused on.

In order to determine what content to dwell on, Journal reporters developed some bots that re-watched videos at random or in a predetermined sequence. Reporters used the results from those bots to generate co-occurrence matrices for each hashtag.

However, because the bots watched many thousands of videos, these matrices were extremely large and computationally challenging to analyze.

To deal with this, reporters used sparse matrices and tools including Truncated SVD to partially reduce the dimensionality of the matrices. They then made a 3D-map of the universe of hashtags

served to their TikTok bots by reducing the resulting vectors down to three dimensions with UMAP, another clustering and dimensionality reduction technique.

In the end, reporters were left with a visualization that placed hashtags that were used in similar ways close to each other and pushed ones used differently far apart. Reporters also clustered the resulting data with DBSCAN to further analyze the results.

For the thumbnail classifications, reporters manually classified several thousand videos as matching or not matching the criteria of interest, such as drug use. They then passed each thumbnail through Google's Inception model (a pre-trained image recognition model), and built multi-layer perceptron, support vector machine, and random forest classifiers on top of the resulting vectors that described each thumbnail.

Finally, in order to find more videos in our set of several hundred thousand that would be interesting to our readers, we built multi-layer perceptron, support vector machine, and random forest classifiers on top of the reduced hashtag vector and the image vector, using manually tagged videos as training data.

**b. How did you verify the results of the method(s) used?**

Journal reporters vetted their results with a data scientist and several algorithm experts. All material that was ultimately published was manually reviewed to ensure the classifications were accurate. In the case of the 3D visualization, reporters spent several days ensuring that the data matched human expectations of hashtag proximity. Furthermore, reporters charted the progress of the bots over the hashtag map, and checked that it matched the bots expected progression.

In the case of the machine learning classifier for images, reporters manually built the training set, and in the training stage, manually checked the test results (holding back 25 percent of the training data for testing). In the case of the models used to surface additional examples, reporters did this to find examples to report out; they did not use it to make broad claims about the dataset writ large. Each example was checked by a reporter.

**c. How did the results of those methods inform your reporting and writing?**

The results led us to confidently report that TikTok's algorithms can send users deep into rabbit holes of content, where almost every video is about a single subject. We found this to be true in a number of categories, including potentially harmful areas like depression, sex and drugs.

The 3D-map generated using UMAP—representing a slice of TikTok's universe—formed the backbone of the Journal's reporting on content served to its various accounts.

Further, reporters quickly discovered that TikTok didn't seem to be differentiating between adults and minors in what content it served. The platform requires new users to enter a date of birth, and the Journal tested a variety of ages for its users. Several 13-year-old accounts were inundated with content inappropriate for minors.

The analysis also revealed that the longer the Journal's accounts stayed on the platform, the more obscure the videos they were served. Through sourcing, reporters were able to determine that these more obscure videos—with lower view counts—were less vetted by the platform's moderators, and therefore more likely to contain problematic content.

**d. Did you seek significant help from social scientists outside your newsroom? Describe their involvement.**

Reporters were fortunate enough to have a professional data scientist on staff who assisted in vetting the project methodology.

**8. Have you run a correction or clarification on the report? Has your reporting received any pushback? Or has anyone come forward to challenge its accuracy or fairness, even if a correction or clarification was not published or aired? If so, please explain the challenge and how it was resolved.**

No.

**9. Difficulty, uniqueness of effort, or other special circumstances related to this subject.**

Work like this had never been done on TikTok. There was no playbook for how to interact programmatically with the app and collect data served to users. To launch their bots, reporters had to learn how to decrypt and decipher the raw HTTPS traffic traveling between the app and TikTok's servers. Reporters then had to find a way to reliably collect, process and act on this data in real time. Finally, reporters needed to make sense of all that information in a way that would be digestible to readers.

All this work had to be done against the backdrop of an app that is technically closed to automation and bulk data collection. TikTok takes active steps to discourage users from collecting and interacting with its platform via bots, and reporters had to regularly overcome new roadblocks introduced by the platform throughout the course of their analysis.

**10. Length of time taken to report, write and edit the story**

Several months.

\*\*\*