Title

Algorithms and the "Anti-Preference": A Quantitative Investigation of “Reaching the Wrong Audience” on TikTok

Journal

Berkeley Undergraduate Journal, 38(1)

Author

Doyle, Owen Alinsangan

Publication Date

2024-11-26

DOI

10.5070/B3.39970

Peer reviewed

ALGORITHMS AND THE "ANTI-PREFERENCE"

A Quantitative Investigation of "Reaching the Wrong Audience" on TikTok

By Owen Alinsangan Doyle

This paper provides an empirical hypothesis test and partial verification of the "algorithmic folk theory" of "reaching the wrong audience." This user folk theory claims that content posted to TikTok can sometimes become sequestered to a hostile audience, resulting in a sharp influx of harassment and oppositional comments. To test this theory, this paper employs a graphical analysis to identify trends in interaction rates, comments over time, and comment sentiment. The data collected consisted of 1,455 posts and 454,540 comments, which were then evaluated using a natural language processing (NLP) sentiment analysis tool for a total of 297,009,882 effective observations of viewer sentiment response. Using this data, this paper employs a time-series analysis to identify a "resurrecting" post behavior characterized by a sudden increase in engagement of an otherwise "dead" TikTok post as far as ten months after the content's initial post date. Further, the findings highlighted how this "resurgent behavior" would commonly occur when the sudden influx of engagement contained either distinctly positive or negative comment sentiment. These findings suggest the existence of "audience sentiment sequestering," explained as the algorithmic restriction of viewership to a specific audience type and a core mechanism of the user folk theory of "reaching the wrong audience." Lastly, this paper proposes a new theoretical algorithmic phenomenon, the anti-preference theory, to explain why automated algorithmic decision-making may cause a user's content to "reach the wrong audience" and remain stuck there. This theory suggests that the recommendation algorithm implemented on TikTok is impartial to the positive or negative sentiment of a viewer's comment but still susceptible to the user's propensity to comment. In conjunction, these traits can cause the recommendation algorithm to "misinterpret" a user's negative comment as a successful recommendation for that viewer. This "misinterpretation" can create a feedback loop, where the recommendation algorithm will show the content to other hostile users with similar "anti-preferences." Expectedly, this hostile audience would share a similar high propensity to leave hostile comments of their own, thus restarting the loop. From the content creator's side, the anti-preference phenomenon can appear as a sudden and seemingly systematic increase in hostile comments, similar to the experiences described within the "reaching the wrong audience" folk theory.

I. Introduction

After its worldwide release in 2018, TikTok quickly became one of the premier social media platforms in the United States and across the globe. TikTok provides users with a unique utilization of short-form vertical video

and a central application of their proprietary recommendation algorithm via the For You Page (FYP). TikTok's advanced recommendation algorithm is most commonly characterized by the way it deweights friends and follower networks and instead maximizes viewer satisfaction.[1] This approach provides a "free-market" social platform approach where "viral-worthy" content can be quickly propelled to the top, regardless of a content creator's existing following.[2] This methodology provided a unique content environment distinct from almost all other social media platforms at the time.

TikTok's explosive growth quickly attracted attention from researchers focused on platform governance and human-algorithm interactions. Most of the research on TikTok mirrors the practices of previous platform research, such as research on the YouTube radicalization pipeline.[3] While this is certainly a valuable perspective within human-algorithm interactions, this approach often focuses entirely on the viewer's experience while overlooking the experience of the content creator. Because research on creator-side harm is still exploratory, one of the best sources for learning about creator issues is "algorithmic folklore." This practice involves platform users, and typically content creators, who create theories that attempt to explain their experience with an algorithm. The "optimistic" practice of algorithmic folklore is best captured by the "algorithmic guru"—someone who shares algorithmic strategies with content creators to help them optimize their content outcomes.[4] The "pessimistic" practice of algorithmic folklore mainly consists of content creators who are attempting to explain any issues they have with the algorithm, such as shadowbanning or content "flopping."[5,6]

One of the most prolific algorithmic folk theories on the TikTok platform is that of "reaching the wrong audience." Broadly, this theory claims that a creator's content can sometimes become sequestered to a hostile audience, resulting in a sharp influx of harassment and oppositional comments. Researchers have already begun to study how content creators on TikTok experience this algorithmic phenomenon. Content creators are recorded as sometimes declaring a post, or even an entire account, as "having ended up on the wrong side of TikTok." Sometimes, creators will even prompt specific and desirable subsets of their TikTok audience to engage with their content as a form of "take back" strategy against the algorithmic decisions. This involves a content creator requesting that people of very specific traits (i.e., young people, LGBTQ+, body positive, alt, etc.) interact with their post to "realign" their audience.[7]

Though extensively recognized within the TikTok content creator community, this algorithmic folk theory has yet to be proven true. To contribute to this end, this paper hopes to empirically verify this folk theory through a quantitative analysis using content metric data extracted from posts across the TikTok platform. Within the paper, the analysis investigates the nature of the TikTok recommendation algorithm and its habits of increasing viewership in "audiences" linked through preference signals, the algorithm's lack of ability to discern between

---

1 Ben Smith, "How TikTok Reads Your Mind," *The New York Times*, December 6, 2021, https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html.

2 Benjamin Guinaudeau, Kevin Munger, and Fabio Votta, "Fifteen Seconds of Fame: TikTok and the Supply Side of Social Video," *Computational Communication Research* 4, no. 2 (October 1, 2022): 463–85, https://doi.org/10.5117/CCR2022.2.004.GUIN.

3 Derek O'Callaghan, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham, "Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems," *Social Science Computer Review* 33, no. 4 (August 1, 2015): 459–78, https://doi.org/10.1177/0894439314555329.

4 Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi, "Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance," *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (October 18, 2021): 305:1–305:44, https://doi.org/10.1145/3476046.

5 Markus Rach, "A Qualitative Study on the Behavioral Impact of TikTok's Platform Mechanics on Economically Driven Content Creators," *International Journal of Social Science and Humanity*, November 4, 2021, 146–50, https://doi.org/10.18178/ijssh.2021.V11.1055.

6 Guinaudeau, Munger, and Votta, "Fifteen Seconds of Fame."

7 Crystal Abidin, "Mapping Internet Celebrity on TikTok: Exploring Attention Economies and Visibility Labours," *Cultural Science Journal* 12, no. 1 (December 31, 2019): 77–103, https://doi.org/10.5334/csci.140.

positive and negative comments, and how these traits can create an algorithmic phenomenon that leaves content creators stuck in a negative feedback loop of hateful comments.

II. Background

II.i. On the Nature of TikTok

The TikTok platform provides a unique opportunity to study recommendation algorithm behavior in the context of social media. Specific design decisions implemented through the platform have opened a new frontier for the applications of social algorithms, such as centering a recommendation algorithm in the form of the For You Page and decreasing the importance placed on networked connections such as followers. As many other social media companies choose to do, TikTok has avoided publishing technical specifics of its algorithm, making the study of the system naturally imprecise.[8] However, using TikTok's specific design affordances, traditional understanding of recommendation algorithms, and the existing information on the TikTok algorithm, important inferences about the structure of the platform's operations can be made.

At its core, TikTok is a short-video-format sharing platform that allows users to view content made by other users across the platform. The most common content on the platform is 3- to 60-second vertical videos, which take up most of a mobile smartphone's screen.[9] Importantly, TikTok centers a recommendation process called the For You Page (FYP) as the main way for users to experience content on the platform. The FYP serves as the platform's landing page and provides users with a continuous stream of content, presented one at a time and navigated with a verticle swipe. Users can choose from several interactions for each video, including liking, commenting, saving, and sharing. One of the defining features of TikTok's recommendation feed is the reduced importance placed on a user's followed accounts. This fundamentally changes the nature of the content users view compared to "networked" social media platforms such as Facebook or Instagram. In the early years of TikTok, this approach was a distinct diversion from other platforms. Soon after TikTok gained popularity, other platforms began to release their own versions of recommendation algorithms, such as Instagram reels and YouTube shorts, both released in 2020. Similarly, across these platforms, the TikTok For You Page creates custom content recommendations based on previous user engagement data. This preference data can be derived from a user's "explicit" interactions of liking and commenting or their "implicit" interactions with the content, such as view time and rewatching. TikTok's recommendation algorithm creates these custom content recommendations to maximize user satisfaction rather than show users the most recent content from their network of followed accounts.[10] This is often considered the main appeal of this modality of social media, creating a unique user experience in which control over a user's feed is partially relinquished to the recommendation algorithm. On the creator side of TikTok, centering the recommendation system places less importance on a creator's explicit following, allowing any creator's content to spread quickly and achieve virality.[11] This can incentivize content creators to remain active on the platform, as any video they post may have the potential for mass virality.

II.ii. Recommendation algorithms

Though the true design of the TikTok algorithm remains unavailable to the public, many have speculated that their system combines collaborative and content-based filtering.[12] Collaborative filtering involves generating a recommendation by comparing the preferences of other users who have acted similarly. Alternatively, content-

---

8 Zoe Ashbridge, "How the TikTok Algorithm Works: Everything You Need to Know," *Search Engine Land*, December 21, 2022, https://searchengineland.com/how-tiktok-algorithm-works-390229.

9 Guinaudeau, Munger, and Votta. "Fifteen Seconds of Fame."

10 Smith, "How TikTok Reads Your Mind."

11 Min Zhang and Yiqun Liu, "A Commentary of TikTok Recommendation Algorithms in MIT Technology Review 2021," *Fundamental Research* 1, no. 6 (November 1, 2021): 846–47, https://doi.org/10.1016/j.fmre.2021.11.015.

12 Zhang and Liu, "A Commentary of TikTok Recommendation Algorithms."

based filtering involves generating a recommendation based on a user's past preferences and producing a similar or related recommendation.[13] Recommendation algorithms consider many different variables when formulating a user's interest signals. For social media platforms, these explicit interest signals are derived from the user's interactions with the content: view time, likes, comments, and shares. In a recent 2023 Ted Talk event, TikTok CEO Shou Chew finally confirmed that TikTok primarily uses content-based filtering, providing the first significant statement from TikTok on the matter.[14] Further, Chew verified that TikTok's algorithm uses the explicit preference signals listed.

Frequently, social recommendation systems come under fire for how they interpret and weigh each of these preference signals. In 2021, *The Washington Post* published an exposé on Facebook's "angry react," using internal Facebook sources to confirm that their algorithm disproportionately favored the angry reaction.[15] This critique of the Facebook recommendation systems aligns well with existing research efforts around feedback loops and rabbit holes. Other research on algorithmic feedback loops typically focuses on how algorithms can influence their users, particularly in ways that skew the user's opinions or worldviews. Research of this nature employs the same understanding of preference signals with a specific focus on identifying echoing or compounding patterns where preferences are amplified through feedback loops.[16] This paper uniquely focuses on the impact of feedback loops on the content creators, rather than viewers. This approach allows for an exploration of algorithmic behaviors on the producing side of TikTok content, one which is often overlooked by researchers.

II.iii. Algorithmic folklore and "reaching the wrong audience"

The centering of virality on TikTok creates a novel set of incentives for content creators. Notably, larger view stratification between viral and non-viral posts can cause increased pressure on content creators to produce viral-worthy content.[17] The pressure to succeed in an algorithmic environment has led to the widespread perpetuation of "algorithmic folklore." These "folk stories" are theories developed by system users that help explain their experience on the platform.[18] Developed first through personal experience, algorithmic folklore is proliferated either through peer-to-peer sharing or dispersion from a "TikTok guru" ("algorithmic guru," or sometimes "TikTok strategist," seen operating within #tiktokgrowth #tiktoktipsandtricks #tiktokalgorithm, and others). These folk theories are meant to help creators gain an "edge-up" in the recommendation processes or otherwise combat the topics of "algorithmic disillusionment" such as shadowbanning or "flopping."[19,20,21]

One of the more widespread folk theories is that of "reaching the wrong audience." This folk theory claims that an algorithmic phenomenon occurs where a TikTok video can become "stuck" in the "wrong audience." This wrong audience is typically considered a viewer demographic—linked through preference, characteristic, or behavior—that is strongly opposed to the affected TikTok video. In accordance with the theory, this phenomenon

---

13 Poonam B.Thorat, R. M. Goudar, and Sunita Barve, "Survey on Collaborative Filtering, Content-Based Filtering and Hybrid Recommendation System," *International Journal of Computer Applications* 110, no. 4 (January 16, 2015): 31–36, https://doi.org/10.5120/19308-0760.

14 Shou Chew, "TikTok's CEO on its future—and what makes its algorithm different," TED2023, April 21, 2023, 06:00, https://www.ted.com/talks/shou_chew_tiktok_s_ceo_on_its_future_and_what_makes_its_algorithm_different/c.

15 Jeremy B. Merrill and Will Oremus, "Five Points for Anger, One for a 'Like': How Facebook's Formula Fostered Rage and Misinformation," *The Washington Post*, October 26, 2021, https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/.

16 Luke Thorburn, "When You Hear 'Filter Bubble,' 'Echo Chamber,' or 'Rabbit Hole' — Think 'Feedback Loop,'" *Understanding Recommenders* (blog), April 4, 2023, https://medium.com/understanding-recommenders/when-you-hear-filter-bubble-echo-chamber-or-rabbit-hole-think-feedback-loop-7d1c8733d5c.

17 Jing Zeng and D. Bondy Valdovinos Kaye, "From Content Moderation to Visibility Moderation: A Case Study of Platform Governance on TikTok," *Policy & Internet* 14, no. 1 (2022): 79–95, https://doi.org/10.1002/poi3.287.

18 Karizat, Delmonaco, Eslami, and Andalibi, "Algorithmic Folk Theories and Identity."

19 Karizat, Delmonaco, Eslami, and Andalibi, "Algorithmic Folk Theories and Identity."

20 Rach, "A Qualitative Study on the Behavioral Impact of TikTok's Platform."

21 Guinaudeau, Munger, and Votta, "Fifteen Seconds of Fame."

is thought to appear as a large and sudden influx of oppositional comments all from a similar demographic. Once a comment section on a specific has become "overrun" by hate commenters, other commenters or sometimes the creator themselves could declare that the post has "reached the wrong audience." Notably, this is one of the only folk theories that both creators and commenters engage with. Across the platform, it is common to see commenters apologizing to an affected content creator in the comments section, leaving comments such as, "Sorry this reached the wrong audience."

In April 2023, this folk theory reached much broader public attention with the publication of a *Business Insider* article that recounted the experience of one specific TikTok creator. The content creator had been running a successful TikTok page that gave travel advice for plus-sized travelers. However, as her content gained popularity, she experienced waves of hateful comments that would entirely overrun her posts.[22] Though this article did not make an explicit connection to the broader folk theory, it still served as an important moment for the public to understand the creator-side harms that can come from social algorithms. In this instance, the content creator had a specific audience in mind, but to her dismay, her content became "stuck" to a hostile audience.

A few social science researchers have already completed important work investigating this folk theory. Most significantly, Riemer and Peter proposed a strong theory around what they refer to as "algorithmic audiencing."[23] They present this phenomenon as the automatic "configuration of audiences" for any given social media content. Applying the technical understanding of algorithms discussed in II.II, this can be identified as a component of collaborative filtering where expressed preferences lead to preference groups. Riemer and Peter argue that this process produces audiences that are deeply segmented by interests and social grouping, contributing to the highly studied "filter bubble" phenomenon. Other notable research includes that of Jones, which begins the important work of exploring how content creators engage with algorithmic audiencing on TikTok.[24] After analyzing the content creators' behaviors, Jones finds that many of them are cognizant that the algorithm shapes the demographics of their audience. This knowledge is clearly shown through a number of behaviors, such as the "If you are seeing this, you are . . ." trend and the "you are now entering" trend. These are trends where content creators will acknowledge the fringe demographic, interest, or preference that resulted in a particular viewer seeing that specific video. By referencing these groupings in their actions—"If you are seeing this, you are in your 20s seeking spiritual freedom," or "you are now entering YA book-Tok," for example—creators are both acknowledging the TikTok algorithm's functions of algorithmic audiencing while also engaging the audiencing function directly. Interestingly, these trends can also be seen as the content creator's attempt to resist any unfavorable algorithmic processes, such as the "reaching the wrong audience" folk theory.

Additionally, researchers have also begun to build theories on how hostile audiences can introduce harm. Jones argues that once a piece of content is exposed to members of a hostile audience, just those few exposures can introduce harassment.[25] Importantly, once these few hostile viewers interact with the content, it becomes part of their "algorithmic audiencing" or "preference grouping." This interaction can then be interpreted by the recommendation system, which will begin to expose more of that hostile audience to the content in question, thus exposing the creator to more harassment.[26] Abidin observes how content creators react to these occurrences, often blaming the algorithm for perpetuating their content to a misaligned audience.[27] Creators will frequently make follow-up posts categorizing previous posts, or sometimes entire creator profiles, as having "ended up on the wrong side of TikTok" or having "reached the wrong audience." Sometimes, creators will even prompt specific

---

22 Andrew Lloyd, "A Plus-Size TikToker Faced a Wave of Abuse after Her Travel Tips for Larger People Went Viral. But the Hate Only Pushed Her to Keep Speaking Up," *Insider*, April 6, 2023, https://www.insider.com/plus-size-tiktoker-abuse-hate-travel-tips-viral-2023-4.

23 Kai Riemer and Sandra Peter, "Algorithmic Audiencing: Why We Need to Rethink Free Speech on Social Media," *Journal of Information Technology* 36, no. 4 (December 1, 2021): 409–26, https://doi.org/10.1177/02683962211013358.

24 Corinne Jones, "How to Train Your Algorithm: The Struggle for Public Control over Private Audience Commodities on Tiktok," *Media, Culture & Society* 45, no. 6 (September 1, 2023): 1192–1209, https://doi.org/10.1177/01634437231159555.

25 Jones, "How to Train Your Algorithm."

26 Jones, "How to Train Your Algorithm."

27 Abidin, "Mapping Internet Celebrity on TikTok."

and desirable subsets of their TikTok audience to engage with their content as a form of "take back" strategy against the algorithmic decisions. This typically involves a content creator declaring that their current audience has become misaligned with their "target audience." They will then request that people of very specific traits (i.e., young people, LGBTQ+, body positive, alt, etc.) interact with their posts. In doing this, the creator hopes that it will change the trajectory of the recommendation system in the future.

II.iv. Research approach

With reference to the technical understanding of recommendation systems from II.ii and the social aspects of this folk theory from II.iii, a formal hypothesis for "reaching the wrong audience" folk theory can be stated. For this to be done, it is integral to draw the connections between technical understandings of algorithms and creator-side folk theory. The first connection is between the technical definition of collaborative filtering and the creator's recounts of "algorithmic audiencing." According to Jones, content creators have been observed acknowledging the audience-sorting capabilities of the algorithm when they call for post interaction from only very specific preference parties.[28] Algorithmic operations such as these can be explained using a technical understanding of recommendation algorithms, and specifically collaborative filtering established in II.ii. Using this definition, the recommendation algorithm uses preference data collected from one user to make the same content recommendation to a different user who has expressed similar preference data. This can mean that the algorithmic audiencing described by creators becomes an expected behavior within a recommendation system that employs collaborative filtering. Further, this means that the development of "algorithmic audiences" could be considered "preference groups" of viewers created when collaborative filtering sorts viewers by their previous history of preference signals. The second important connection between technical understandings and creator experiences is the relationship between the creator's experience of hostile comments and the nature of recommendation algorithms. Within the folk theory of "reaching the wrong audience," creators describe a situation where negative comments propel exposure to more viewers who will also leave negative comments. Using the technical understanding of recommendation algorithms established in II.ii, the plausible explanation is that the TikTok algorithm cannot understand the sentiment of each comment. This would mean that the recommendation algorithm is unable to differentiate if a viewer leaving a comment should be considered a successful recommendation or an unsuccessful recommendation. Instead, the recommendation algorithm will interpret a viewer leaving a comment as a successful recommendation regardless of the nature of the comment. This oversight allows a hostile comment to be interpreted as a successful recommendation, signaling the algorithm's collaborative filtering processes to show the same content to similar audiences who may also have a heightened propensity to leave hostile comments and thus reinforce the recommendation.

> Hypothesis: Creator-made content on TikTok can become "stuck" to hostile audiences through an algorithmic shortcoming where a viewer's negative comment is interpreted as a successful recommendation. This causes the content to be shown to an increasing number of hostile viewers with a similar high propensity to leave a negative or hateful comment. This results in a negative feedback loop where content creators are barraged with hostile comments.
> P1: TikTok's recommendation system operates in "expanding audiences."
> P2: A creator's desired or undesired content outcome can be linked to the sentiment of comments.
> P3: The undesired content outcome occurs through an algorithmic process that expands the content's audience based on a viewer's propensity to comment and has an observable effect on comment sentiment.

28 Jones, "How to Train Your Algorithm."

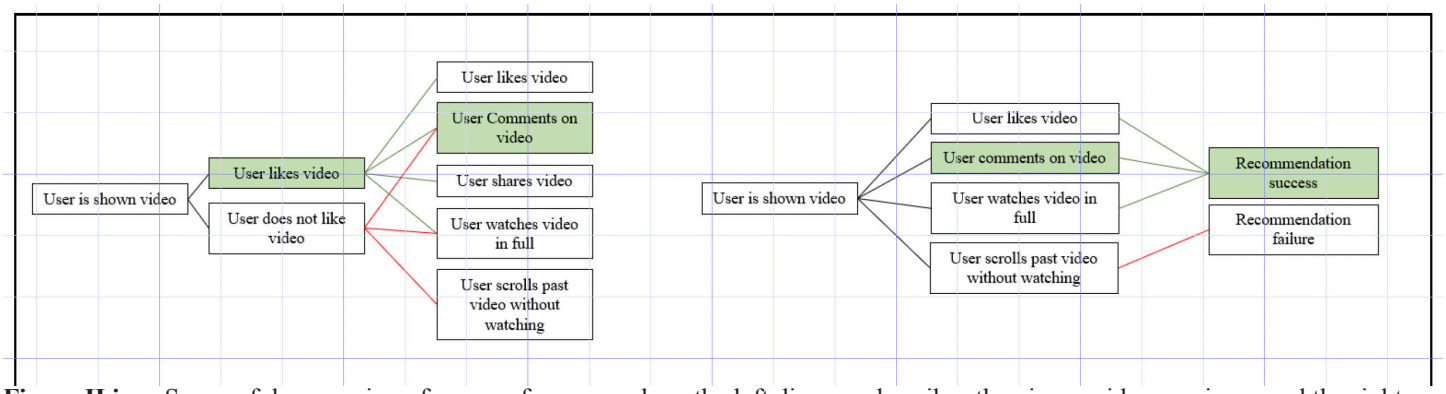


Figure II.iv.a. Successful expression of user preferences, where the left diagram describes the viewer-side experience and the right diagram describes the recommendation system's interpretation. In this case, a viewer comments because they enjoyed the content. The algorithm accurately interprets this as a recommendation success, as seen in the right diagram.

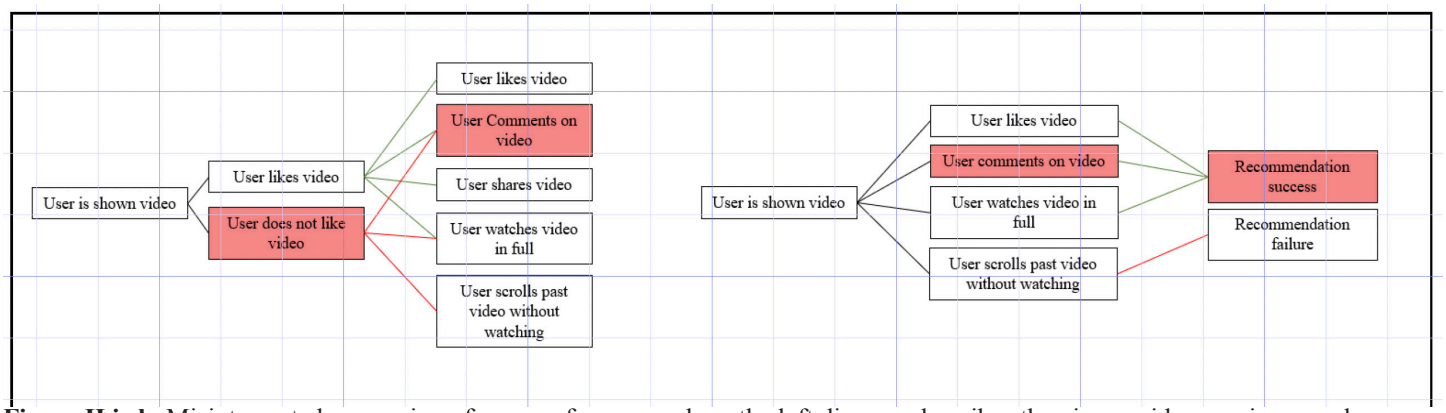


Figure II.iv.b. Misinterpreted expression of user preferences, where the left diagram describes the viewer-side experience and the right diagram describes the recommendation system's interpretation. Because users will still comment when they don't like a video, this can cause a misinterpretation by the recommendation system, as seen in the right diagram. In this case, a user comments something hostile because they don't like the content they were shown. Because the algorithm cannot delineate between positive and negative comments, the algorithm will interpret the given preference signal as a recommendation success. Because of this, the algorithm will show the same video to similar viewers.

III. Methodology

III.i. Data collection

III.i.a. TikTok API

Until early 2023, TikTok did not offer access to an extensive application programming interface (API), which would allow developers and researchers to interact with the publically-visible data on the platform. Since its U.S. release in 2018, TikTok has offered only a "Display API" through its "TikTok for Developers" website. This simple HTTP-based API allows developers to view a TikTok page's public information in the form of text. Compared to the comprehensive APIs available on social media platforms such as Twitter (now known as X) and YouTube, TikTok's Display API remains underwhelming and provides significantly less research functionality. However, as of February 21, 2023, TikTok released a separate "Research API" available to researchers upon application and approval. The Research API allows greater access to platform data, including public user profile information such as post characteristics, the number of post likes, and the number of comments. Further, the Research API allows an expanded version of the platform's existing keyword search functionality. While this would have benefitted this project, the TikTok Research API was not used because of the considerable approval time needed to gain access. Instead, this project uses a common technique called data scraping, which is often used to collect data for quantitative studies when a platform's API is not available.

III.i.b. Data scraper and NLP sentiment analysis

Instead of the TikTok Research API, this project used a common third-party data scraper. This tool operates by accessing the HTTP-based Display API and using brute-force automation to extract specific publicly available data from the TikTok platform. The resulting outputs resemble those that would otherwise be accessed through the TikTok Research API: publicly available user profile statistics, including post characteristics such as the number of post likes and the number of post comments. The scraper implemented consisted of two separate sets of code. The first script allows the researcher to scrape posts from target hashtags by inputting a target hashtag and receiving an output of the top posts under that hashtag. The second script allows the researcher to scrape a specific post's comments by inputting a direct link to a post and receiving an output of comments left on the specified post. Data scraping on social media platforms is recognized as fair use, and TikTok only partially restricts third-party scrapers, primarily focusing on preventing mass data scraping efforts. All scraped data in the project is publicly available on the TikTok platform and could otherwise be accessed by the average TikTok user. All scraped data was held on a secure network, and no user-specific traits, such as political opinions or countries of origin, were direct factors in this analysis. For the application of this paper's results, all user information is anonymized, and measures have been taken to prevent back-tracing. This paper also employs large-scale sentiment analysis techniques. For this, a popular third-party sentiment analysis tool was used. This tool uses a pre-trained data categorization model that employs Natural Language Processing (NLP) techniques, including the AI model's ability to interpret human language. The model used in this project was trained on a significant index of Twitter comments and calibrated for internet slang, irony, and other irregular language use seen on Twitter. This model uses inputs of comment text for an output of a sentiment score with a range of -1.0 to +1.0. Posts ranging from -1.0 to -0.26 are considered negative, -0.25 to 0.25 are considered neutral, and 0.26 to 1.0 are considered positive.

III.ii. Sampling and analysis

III.ii.a. Sampling approach

Due to the large volume of content on TikTok, this project employs non-probability sampling techniques. The main sampling method used was non-proportional quota sampling, where a total population is divided into characteristic groups, and sub-samples are taken from those groups in proportion. For this project, the total population was determined to be highly viral content on TikTok. From this population, three content groups were selected based on a pre-established understanding of their characteristics: #politics, #PetsOfTikTok, and #based (see Table III. ii.a). These strata were selected to represent a spread of content characteristics and the expected characteristics of their interactions.

After establishing the strata, a structured sample is taken from each hashtag. For this paper's interaction rate analysis section, the top 947 posts from each hashtag are sampled. This represents the top 947 most viral posts under the given hashtag, ranked by the total number of likes per post. Roughly 947 is the maximum number of posts TikTok will display under any given hashtag. For the sentiment analysis section of this paper, the top 485 posts from each hashtag are sampled. This smaller sample size was necessary due to computing cost limitations.

For the sentiment analysis section of this paper, a structured sub-sample of comments was needed. After identifying the top 485 posts from each hashtag, the top 350 comments from each post were collected. Typically, the TikTok mobile and web applications will not load more than ~350 comments per post. These comments are displayed in a pre-set order, understood as a combination of all-time top-liked comments and "algorithmically hot" recent comments. If the total number of comments selected from these two categories is below 350, then TikTok will also display the most recent comments, which often have fewer likes.

Overall, this paper's structured quota sampling process sacrifices the generalizability of its findings. This prevents this project from representing the holistic landscape of TikTok posts, specifically, the significant volume of non-viral TikTok content on the platform. A generalizable version of this project would focus on a stratified sample. This would be done by randomly selecting many hashtags on the platform, surveying the entirety of their posts, and using random or systematic sampling to select a large number of posts for analysis. Considering

these requirements, this method would be incredibly resource-intensive and may only be possible with privileged access to internal TikTok databases that is only granted to internal TikTok researchers.

Table III.ii.a. Breakdown of content groups, their expected traits, and typical viewer interactions.

	Education	**Entertainment**	**Self-help & comedy**
	#Politics	#PetsOfTikTok	#Based
Content Characteristics:	Contentious, political, discussion-based, high-engagement	Non-contentious, positive, affirmative, agreement.	Apolitical, contentious, lifestyle, one-sided, affirmative.
Interaction Type:	Discourse, arguments, intense agreement or disagreement.	Agreement and exclamations regarding the subject matter of the video.	Viewers intake lifestyle opinions or memes and affirm content relevancy or humor.
Content Intentions:	Serious subject matter that is debate-driven. Polarized sentiments.	Universally positive subject matter. Positive sentiment.	Content typically has a pessimistic worldview. Negative sentiment.
Target Audience:	Politically-involved users.	Consumers of "general content," pet owners.	Young men, people involved with internet and meme culture.
Tags and Themes:	News, debate, current issues, Republican, Democrat, elections, policy, taxes, and economy.	Cats, dogs, cute, funny, relaxing, uplifting.	Lifestyle, internet culture, male discourse, masculinity, success, far-left and far-right values (apolitical), memes, stoicism.

III.ii.b. Building and cleaning datasets

Sampling occurred in April 2023. The first dataset (used for section IV.i) contained 2,842 total posts collected across the three target hashtags. The initial goal of 3,000 total posts was unmet because the TikTok platform limits the maximum number of viewable posts under each hashtag. At the time this data was collected, the average maximum number of viewable posts was 947 per hashtag.

The first dataset contains data across 3.4 billion likes and 22.2 billion views. The selection of videos makes up roughly 11 percent of the total views of #politics, 19.3 percent of the total views of #PetsOfTikTok, and 24.6 percent of the total views of #based. This dataset includes all the significant variables on a post's performance: *Total Views*, *Total Likes*, *Total Comments*, *Total Shares*, and *Hashtags Used*. From here, interaction rates can be calculated using the target interaction divided by *Total Views* to return the additional variables *Like Rate*, *Comment Rate*, and *Share Rate*.

An additional two datasets (dataset IV.ii and dataset IV.iii) were developed for the sentiment analysis sections of this project, IV.ii and IV.iii. The process began by first creating the dataset for IV.iii. This project first selected the top 500 posts from each selected hashtag. This number was reduced from the initial 947 seen in dataset IV.i due to computing costs. Because the creation of this dataset occurred over several days, a significant number of posts selected would either drop from the hashtag, be removed by TikTok, or be removed by the creator. By the end of creating dataset IV.iii, the dataset was normalized to contain 485 posts per hashtag for a total of 1,455 posts. Using these posts, this project employed a TikTok data scraper to collect the top 350 comments from each post's comment section, resulting in roughly 509,600 comments surveyed. For use in text-based sentiment analysis, this dataset was cleaned of all comments that only contained non-character elements, such as emojis and character symbols. This resulted in 454,540 total comments. This dataset contained key variables such as *Comment Text*, *Time Of Comment*, and *Total Comment Likes*. After this data was collected, the *Comment Text* for each comment

was analyzed using NLP sentiment analysis. This returned additional information about each comment: *Sentiment Score*, *Sentiment Categorization*, *Sentiment Occurrence*, and *Text Themes*.

Using dataset IV.iii described above, dataset IV.ii was created with an additional calculation. Using the *Sentiment Score* and *Total Comment Likes* variables of each comment, an additional variable *Weighted Sentiment* was calculated. This was calculated by multiplying (*Total Likes* + 1) by the *Sentiment Score*, where the "+1" represents the original commenter "agreeing with" or "voting for" their own comment. Importantly, this technique counts any user's "liking" of a comment as a "vote" for that comment's *Sentiment Score*. By creating the *Weighted Sentiment* variable, the sentiment analysis data can be expanded for each comment section. On a per-post basis, this can expand a post's sentiment information from 350 sentiment data points—one for each comment collected per post—to an average of 198,006 sentiment data points per post. As a whole, this calculation expanded the total points of sentiment information from 454,540 to a total of 297,009,882 effective observations of viewer sentiment response. After the weighted sentiment calculation, the *Weighted Sentiment* was averaged across 385 posts from each content group. This created dataset IV.ii, which consisted of the top 485 posts from each hashtag, their standard variables such as *Total Views* and *Total Likes*, and the newly calculated *Average Sentiment* of each post.

IV. Findings

The findings section presents an empirical hypothesis test of "reaching the wrong audience" within TikTok's recommendation system. The section adopts a three-part approach that covers the three core premises of the hypothesis. Within each sub-section, the analysis attempts to present data that can corroborate each premise and thus test the validity of the hypothesis as a whole.

> Hypothesis: Creator-made content on TikTok can become "stuck" to hostile audiences through an algorithmic shortcoming where a viewer's negative comment is interpreted as a successful recommendation. This causes the content to be shown to an increasing number of hostile viewers with a similar high propensity to leave a negative or hateful comment. This results in a negative feedback loop where content creators are barraged with hostile comments.
> P1: TikTok's recommendation system operates in "expanding audiences."
> P2: A creator's desired or undesired content outcome can be linked to the sentiment of comments.
> P3: The undesired content outcome occurs through an algorithmic process that expands the content's audience based on a viewer's propensity to comment and has an observable effect on comment sentiment.

The first section of this analysis (IV.i) investigates how viral content behaves as it becomes exposed to greater volumes of audiences. Here, a cross-sectional analysis is used to observe how content performs as viewership increases. This analysis finds reasonable evidence that TikTok's algorithm expands in conceptual "audiences" or interest groups by testing a video against new viewers until interest groups are exhausted. The second section (IV. ii) introduces sentiment analysis as a new measure of viewer feedback. This employs a similar cross-sectional visualization that measures comment sentiment as *Total Views*, *Like Rate*, and *Comment Rate* increase. This section establishes a framework for approaching sentiment analysis within the context of a recommendation system. Using the same sentiment analysis framework, this analysis proposes a more accurate method for defining a "distinctly undesirable" or "distinctly desirable" post outcome from the point of view of the content creator. The third and final section (IV.iii) combines the sentiment analysis system and a time series analysis to evaluate the way comments change over time on an individual post. This analysis uses sentiment scores collected from a post's comment section to graphically describe how the recommendation system creates "resurgent" video behaviors. The analysis finds that sometimes a post can become "resurrected" months after its initial algorithmic popularity and is also sometimes sequestered to particular sentiment reactions from viewers, partially verifying the hypothesis of this paper.

IV.i. Interaction rates as explicit viewer feedback

This section focuses on testing P1: "TikTok's recommendation system operates in 'expanding audiences.'" This premise argues that the TikTok recommendation system categorizes viewers into "algorithmic audiences." This concept describes how TikTok's recommendation system efficiently spreads content to increasingly large viewership based on the input signals of the previous set of viewers. Algorithmic audiences are thought to be created by grouping viewers with similar interests and preferences and spreading content within those groups. Premise 1 argues that content can be introduced to an algorithmic audience through a process of marginal exposure, where content is tested on fringe viewers and their interactions measure recommendation success. If testing on specific viewers is successful, the content is then shown to an expanding number of viewers within that specific algorithmic audience. These assumptions are consistent with this paper's understanding of collaborative filtering but still necessitate empirical backing.

This section employs the tools of simple interaction rates to prove the application of collaborative filtering as a form of algorithmic audiences. Simple interaction rates such as like rate and comment rate are interpreted as ways in which viewers *explicitly* communicate their preferences to the recommendation system (as opposed to the *implicit* communication of comment sentiment, explored in IV.ii and IV.iii). The data used for this section consists of the total views, likes, comments, and shares collected from the top 485 most-liked videos across the #politics, #PetsOfTikTok, and #based hashtags, for a total of 1,455 posts analyzed. Additional metrics of like rate, comment rate, and share rate are calculated by taking the occurrence count of the specific interaction and dividing it by the total views of the video. Using these calculated metrics combined with the interpretation of interactions as *explicit* communication with the recommendation system, the following tables can be deduced.

Table IV.i.a. Viewer actions are defined as feedback given to the recommendation system.

Viewer action:	Viewer feedback being signaled:
View video	Every viewer that comes across the video on TikTok's FYP will add a view to the total view count.
Scroll past	Viewer identifies that they do not like the content and moves on without interacting. Watch time is an evaluation metric for TikTok, but it is unidentifiable in the context of this project.
Like	Viewer identifies that they like the content, so they tap the like button.
Comment	Viewer identifies that they either like/agree with the content, or that they do not like the content and would like to dispute it.
Share	Viewer identifies that a friend may like or dislike the content.

This table describes the various ways in which viewers interact with content and why they do so. These are considered *explicit* feedback actions because they involve a direct numerical value. The second column lists the viewer's reasoning for executing the viewer's action. It should be noted that TikTok's recommendation system cannot always accurately capture this intent.

Table IV.i.b. Defining the variables used in the analysis.

As metric increases:	**It measures:**
Total views	Increasing exposure to general audiences. This includes aligned and misaligned audiences.
Total likes	Increase exposure to an aligned audience.
Total comments	Increasing exposure to a general audience, scaling with total views and likes. This includes aligned and misaligned audiences.

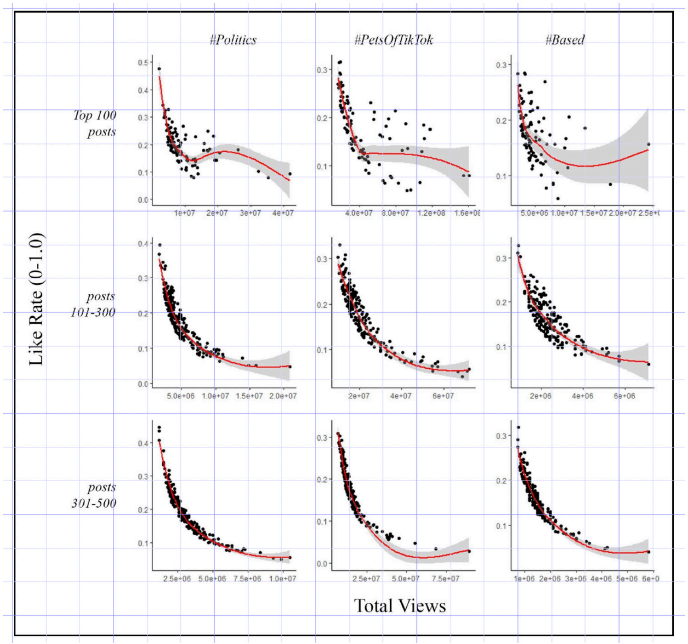


Figure IV.i.a. Cross-sectional analysis of total views against like rate, where each plotted data point represents a single video collected from the hashtag. Post groupings 101–300 and 301–500 display clear trends of decreasing like rate as total views increase, meaning a decreasing portion of people like the content as viewership expands. This implies that when posts are exposed to greater audiences, a post experiences more *audience exposure failure*. In the instance of a failed recommendation, a viewer does not like the post and only scrolls past, adding to the total view count while not adding to the total like count, thus decreasing the like rate.

Variables that represent content characteristics:	
Like rate	Serves as the main success metric for given content. A high like rate signals natural content likability or success in reaching the aligned audiences.
Comment rate	Serves as the main controversy metric for given content. A high comment rate signals either high agreeability or high controversy, which are both effected by the alignment of the audience.

This table interprets the viewer actions from Table IV.i.a into measurable variables for graphical analysis. The left column describes the variable, and the right column describes its meaning. The first two rows are variables that focus on algorithmic performance, as they scale with exposure and do not internalize the features of the post itself. The last two rows are variables that focus on the features of the post. These variables are derived from the first three metrics and are used to describe both the natural features of the post as well as the context of its algorithmic performance. Because this second category of variables also factors in non-interaction rates, they are far more accurate descriptors. In section IV.ii, a new variable *Sentiment Score* will be introduced. This variable would similarly fall into the second category of variables.

IV.i.a. Results

To understand Figures IV.i.a and IV.i.b, it is important to know that most of the videos surveyed are no longer “algorithmically active.” This refers to the videos being in a state where they are no longer accruing significant views and have lost their “algorithmic momentum.” This is typically caused by a video becoming a certain age (anywhere from 1 to 12 months) where it is no longer favored by the recommendation system.

Overall, Figure IV.i.a makes clear a significant trend in viral content. As seen in post groupings 101–300 and 301–500, when a post is exposed to greater audiences, it experiences a higher portion of non-interactions

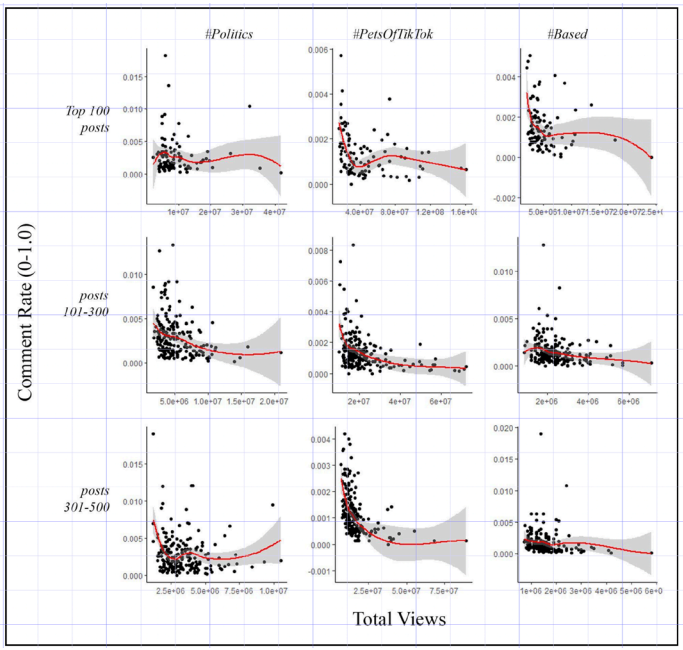


Figure IV.i.b. Cross-sectional analysis of total views against comment rate, where each plotted data point represents a single post collected from the hashtag. In line with the theory developed from Figure IV.i.a., there is far less correlation between increasing views and changes in the likelihood of a viewer to comment. This difference can be explained by viewers choosing to comment either when they like the content *or* when they disagree with the content. This varies from the "like" interaction of Figure 4.1a and causes greater variation in the comment rate trend as total views increase.

expressed by viewing a video without liking it. These non-interactions can be considered recommendation failures, where the recommendation system tests exposing the content to an additional audience. Videos with lower total views, seen on the left portions of each graph, are shown to have maintained tighter interest group exposure—and thus a higher like rate—before falling out of virality over time. Alternately, videos with higher total views, seen on the right portions of each graph, show looser exposure to interest groups during a similar period, seen through the lower like rates and higher percentage of non-interactions.

This content behavior is likely caused by the relationship between the content's intrinsic properties and the mechanisms of the recommendation algorithm. A content's intrinsic properties reflect its natural "likeability," which can be considered the potential for a specific like rate. When injected into the recommendation system, the post has a fixed amount of time to achieve a like rate. During that timeline, the recommendation system will expand viewership by showing the content to an increasing number of viewers. This opens the content to exposure of non-interactions when it is shown to audiences who do not like the content. After achieving high saturation levels in a content's preferred audiences, most posts will start to see a decrease in like rate as they become exposed to more non-interaction groups, as shown in Figure IV.i.a. However, videos with a higher natural likeability or that appeal to wider audiences are seen to have a higher resistance to exposure to new audiences. These videos can maintain comparatively high like rates while also achieving high view counts, causing them to enter the top-100 grouping seen in Figure IV.i.a.

Regarding P1 testing, Figure IV.i.a provides the first layer of empirical evidence to show that TikTok's recommendation system does expand viewership by interest groups. As discussed in the previous paragraphs, the data presented shows a decreasing interest in a post as its viewership increases. This trend implies that the recommendation system frequently tests the limits of positive viewer preference signals as a part of viewership expansion. This finding, combined with this paper's understanding of collaborative filtering, can verify the existence of algorithmic audience expansions within TikTok's recommendation system. Important for sections IV.ii and IV.iii, this finding also suggests that highly viral content is more likely to be exposed to fringe audiences, which can also increase the chance of exposure to a hostile audience.

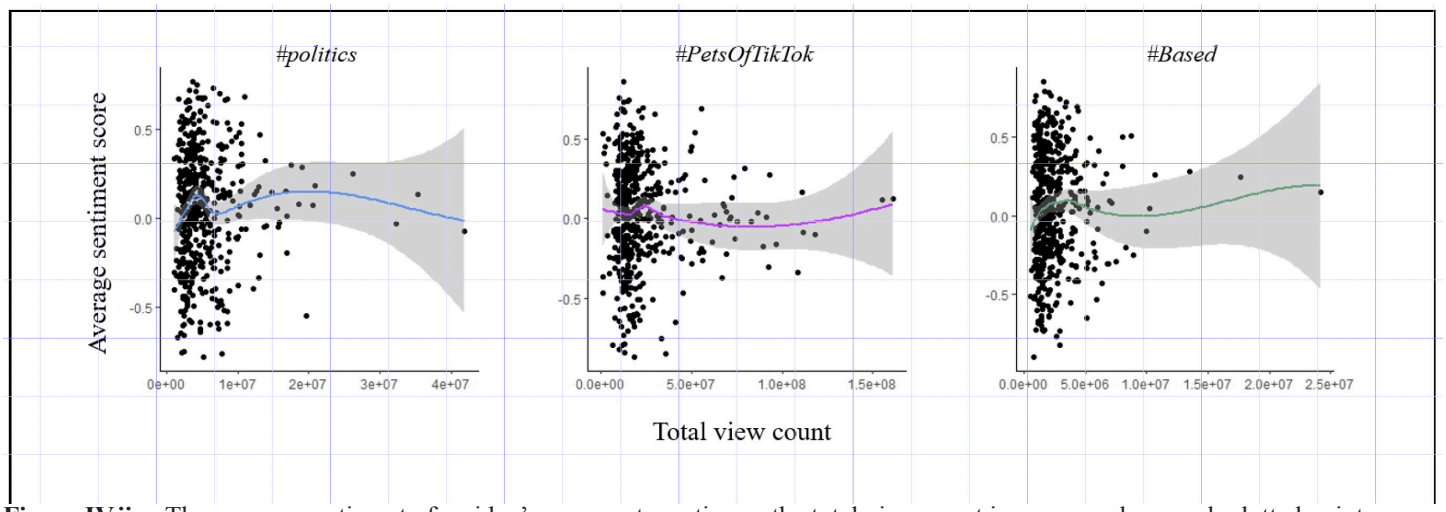


Figure IV.ii.a. The average sentiment of a video's comments section as the total view count increases, where each plotted point represents a single video collected from the given hashtag. This figure shows an even distribution across the sentiment levels for all ranges of view counts. Little to no variation in sentiment is recognizable across the content types. This suggests that sentiment is independent of content type and more reliant on the specific intentions of each video.

IV.ii. Sentiment analysis as implicit viewer feedback

This section introduces sentiment analysis as a novel way to measure viewer feedback on TikTok. This section focuses on testing P2: "A creator's desired or undesired content outcome can be linked to the sentiment of comments." Figures IV.ii.a, IV.ii.c, and IV.ii.d focus on developing a framework for defining a "desired or undesired content outcome from a creator's perspective." Specifically, these figures focus on the relationship between content failure, content success, and sentiment analysis. This framework is described in Table IV.ii.b and then used in the following section IV.iii: sentiment inflection points of individual posts.

IV.ii.a Results

For Figure IV.ii.a, the total view count was chosen as the X-axis because it encompasses a video's exposure to both positive and negative audiences. This provides the most accurate characterization of sentiment trends, as other metrics such as the like rate seen in Figure IV.ii.d introduce heightened levels of confoundment. Further, it is important to recognize that because of the averaging method used, videos with an average sentiment score are not necessarily neutral. Instead, a near-zero average sentiment score can signal that the video had neutral sentiments or otherwise was highly and evenly contested with equally polarized reactions. This is considered when the selection methodology for Section IV.iii is determined later in this paper.

With this in mind, Figure IV.ii.a presents a significant finding for the application of sentiment analysis in this paper. Across all content types, average sentiment scores demonstrate a highly even spread through the full Y-axis range. The similarity between the three highly divergent content styles, as well as the even spread of sentiment through the core viewership range, imply two important points. First, that sentiment is independent of content type and is therefore more reliant on the specific intentions of each video. Secondly, sentiment extremes are more easily achieved in lower-viewership settings, as fewer sentiment extremes are present as views increase. Using these points, the analysis can address the objective of defining an undesired outcome from a content creator's perspective, according to Premise 2. Seeing that any given video can exist at either sentiment extreme and still be considered a "successful content outcome" by the content creator, a definition for an undesired content outcome cannot be linked to negative sentiments. Instead, the definition must be expanded to include the content creator's intentions. By expanding the definition to include intent, it now considers that content creators can make content with intended positive or negative sentiment outcomes across all three content types, as demonstrated in Figure IV.ii.a. Thus, it can be concluded that a "negative content outcome" from the perspective of a content creator occurs when there is an exceptional presence of the sentiment that is opposite from what was intended.

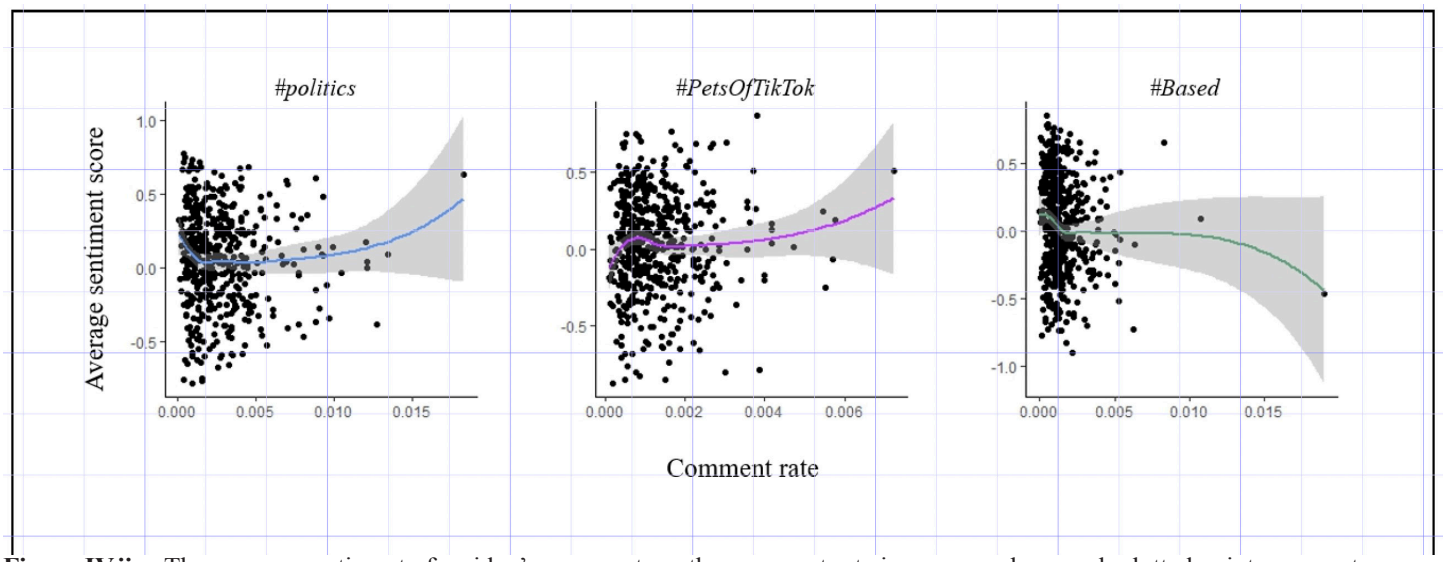


Figure IV.ii.c. The average sentiment of a video's comments as the comment rate increases, where each plotted point represents a single video collected from the given hashtag. This figure shows a similar even distribution of sentiment scores across the full range, as in Figure IV.ii.a.

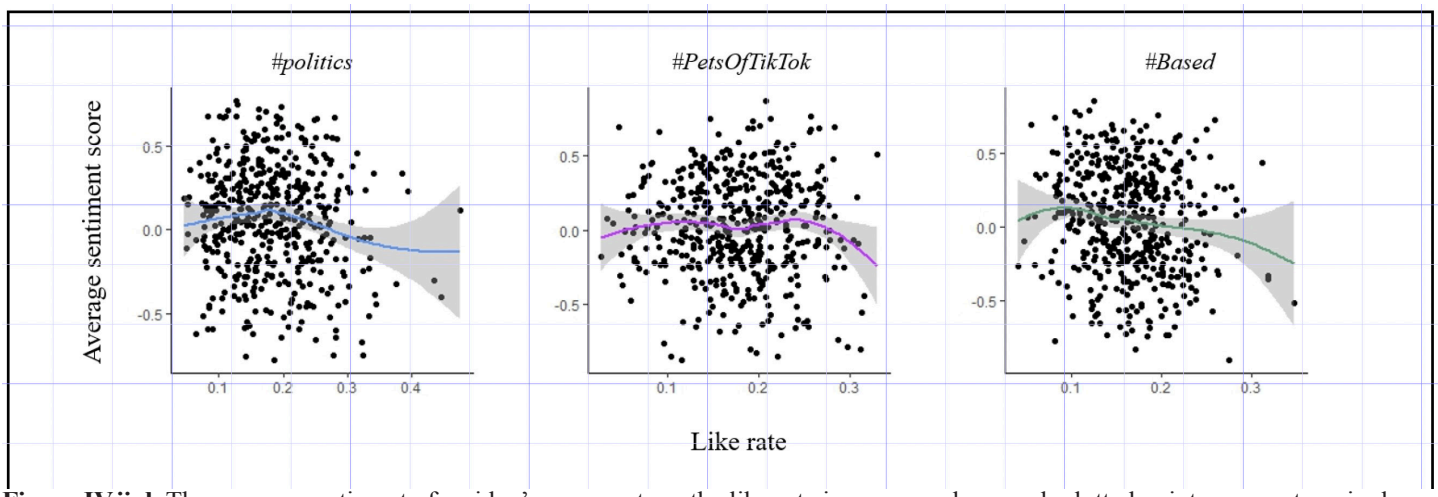


Figure IV.ii.d. The average sentiment of a video's comments as the like rate increases, where each plotted point represents a single video collected from the given hashtag. This figure shows an even distribution of sentiment scores across the full Y range and across the like rate range as well. The lack of clustering as the like rate increases further implies that the sentiment score is not correlated to desired or undesired content outcomes. This provides additional evidence for the intent-based definition highlighted in Table IV.ii.b.

Table IV.ii.b. Definition of successful and unsuccessful post outcomes from the content creator's perspective.

	Positive content	**Negative content**
Positive sentiment reaction	Positive content, positive sentiment reaction	Negative content, positive sentiment reaction
Negative sentiment reaction	Positive content, negative sentiment reaction	Negative content, negative sentiment reaction

This table describes how sentiment scores and post intentions interact to define the success of post outcomes, where the green color signals desired content outcome and the red color signals undesired content outcome.

IV.iii. Resurgent behaviors and sentiment inflection points

The final section of this analysis tests hypothesis P3: "The undesired content outcome occurs through an algorithmic process that expands the content's audience based on a viewer's propensity to comment and has an observable effect on comment sentiment." This section builds on the findings presented in the last two sections to propose a

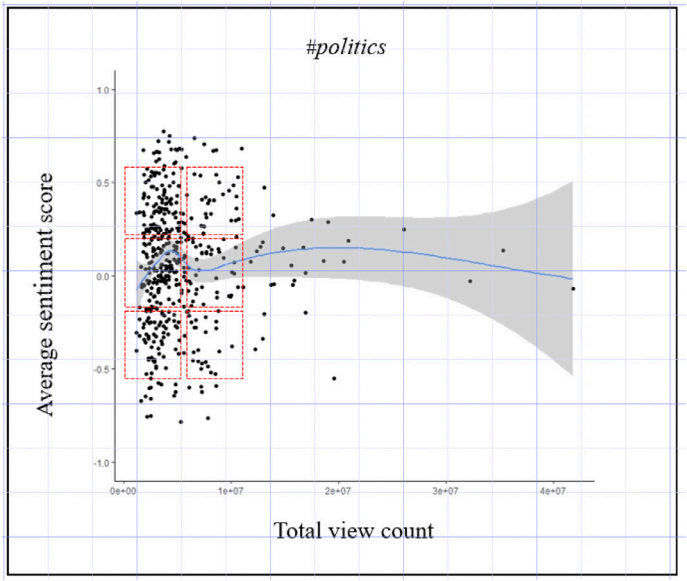


Figure IV.iii.a. Illustration of stratified sampling methodology used for #Politics. Strata were established using range limitations on the X and Y axes. Data points were then selected randomly within each stratum. This was repeated for each content category.

new way to graphically test for a post "reaching the wrong audience." This test is done with a post-to-post analysis where the date of each comment is plotted against the sentiment score that the comment received. Considering analysis 4.1 shows that the recommendation system expands in audiences, it is expected that sentiment trends will cluster to certain sentient score levels. According to the theory of audiences, each sentiment cluster can serve as a proxy for graphically viewing an interest group. However, without introducing a qualitative evaluation of each post analyzed, the intended sentiment of the original post cannot be determined. To work around this, this test focuses on *sentiment inflection points*: a point in time after a video's initial comment sentiment is established where the average sentiment of the comments suddenly switches polarity. This allows the analysis to assume that the initial sentiment seen was "naturally" occurring and representative of a desired post outcome and that the polarity switch represents the point where the post became "stuck" in the wrong audience. The test searches for polarity switches that occur outside of the natural or expected lifetime behaviors of a given post. This will appear graphically as a large influx of opposite-polarity comments many days or weeks after the initial post momentum had supposedly declined. Unnatural content behaviors like this suggest an algorithmic process was involved in the post's "resurrection." If this were true, it would also mean that any polarity switch—and subsequent sequestering of comment sentiments to that specific polarity—can also be attributed to an algorithmic process.

With this testing methodology established, this analysis selected six posts from each content type for a preliminary test. This set of 18 total posts was chosen because of its even distribution across each content type. Specifically, each of the six posts represents a target selection criteria:

Post 1: Low total view count, negative sentiment.
Post 2: Low total view count, neutral sentiment.
Post 3: Low total view count, positive sentiment.
Post 4: High total view count, negative sentiment.
Post 5: High total view count, neutral sentiment.
Post 6: High total view count, positive sentiment.

Each post was selected using a stratified random sampling methodology. This was completed by placing a filter on the dataset to restrict the available data to only the target view count and sentiment zone. Then, a random number generator was used to select a post from the target zone.

IV.iii.a Results

After observing the post selections, the behaviors were placed into three categories: expected behavior, resurgent behavior, and inflection behavior. The expected behavior classification comprised posts that acted " naturally." This involves a front-loaded spike in comment volume immediately after the origin date (initial post date). This is accompanied by an equal distribution of sentiments across the three sentiment categories, though it does not necessitate it. Importantly, the expected behavior class sees a gradual and maintained downward trend in comment volume after the initial comment load. This is expected from the recommendation system, as time elapsed from the origin date is known to be a force of decay for comment volume. The second behavior group that was identified was resurgent behavior. This classification comprised posts that experienced an anomalous resurgence in comment volume a significant time after the initial comment load ended. Typically, these posts first appear the same as the expected behavior classification, displaying a front-loaded spike in comment volume immediately after the origin date. Also similar to the expected behavior group, this initial comment load is typically equally distributed across the sentiment range. However, these posts uniquely saw an additional influx of comments days, weeks, or months after the initial comment load had decayed. The resurgent comments of this behavior group demonstrate similar equal distribution across the sentiment range, indicating no significant change in sentiment upon resurgence. The existence of this resurgent behavior has significant implications for the hypothesis, as it provides evidence that the recommendation system can "revive" a post given the correct conditions. To fully test the hypothesis, the third grouping—inflection behavior—must also be considered. According to Premise 3 of the hypothesis, the inflection behavior group must exemplify an algorithmic force that pushes a post toward an undesirable sentiment outcome. A similar trend is observed in this analysis, though with some key differences. The observed inflection behavior group appears similar to the resurgent behavior group, with both groups starting with a high comment volume and a subsequent volume decline, followed by a resurgence in volume at a later date. However, what differentiates the two groups can be found in the nature of the inflection group's resurgent comments. In both observed instances of inflection behavior, the resurgent group saw a disproportionate number of *positive* sentiment comments. After completing content analysis on the selected posts, it appeared that the positive sentiment comments were considered a content success (see Table IV.ii.b.). This fails to fully match hypothesis P3 because while disproportionate resurgence sentiments were identified, they do not exert harm on the creator as hypothesized. Instead, within the group observed, this second wave of comments brings further post success with reduced exposure to misaligned audiences. With such sequestered sentiments focused on a favorable sentiment group for that post, the inflection behaviors observed only bring the creator increased post success.

In completing this analysis, a number of other significant findings were uncovered. Specifically, the analysis found interesting initial correlations between content characteristics and the propensity to resurge. Most shockingly, what the analysis first explained as the "expected behavior" may not be the most common. Of the 18 posts observed using probability sampling, only seven (38.9 percent) were of the expected behavior classification. Instead, the dominant group proved to be resurgent behaviors, observed at nine total posts or 50 percent of the sample. If inflection behaviors (2 posts, 11.1 percent) are temporarily included as a technical member of the resurgent group, then the resurgent number increases to 61.1 percent. This would suggest that any creator within the top 485 posts of a given hashtag could reasonably expect some sort of resurgent wave of engagement on their post, ranging as far as ten months after its initial post date. Unfortunately, these findings cannot be generalized to the broader content climate on TikTok because the sampling methodology only selects from already highly viral posts. This, however, opens up a unique opportunity for a large-scale study to approach virality research from this novel perspective.

Lastly, additional trends were identified that suggest a correlation between viewership, sentiments, and resurgent outcomes. Across the seven posts identified as expected behavior, 5 of them (71.4 percent) were from posts that contained the "low view" or "negative sentiment" post trait (Table 4.3b.). Interestingly, resurgent behaviors seemed to favor posts with neutral or positive sentiment averages, counting for 8 of 9 total posts (88.9 percent) or 10 out of 11 (91 percent), if including the inflection group. Overall, the evidence shows that the recommendation system does not like to let posts "die" and will instead constantly seek new audience expansions. Unfortunately, the data cannot accurately identify the causal reasoning behind these expansions beyond simply

proving their existence. Considering the correlations identified, the data roughly suggests that high-view, neutral-to-positive posts have a significantly higher chance of having resurgent outcomes. This point is only reinforced by the fact that both of the inflection behavior occurrences came from high-view, positive sentiment posts. Importantly, the data cannot make these causal connections because the resurgent expansion confounds both the total views variable as well as the sentient variable. That being said, this research may serve as a good starting point for a more comprehensive statistical investigation where sampling can be expanded beyond hashtags and advanced control variables can be implemented.

Table IV.iii.b. Observed occurrence rates of three behavior categories, divided by trait.

	Low views	**High views**	**Negative sentiment**	**Neutral sentiment**	**Positive sentiment**	**Total**
Expected	5	2	5	1	1	7
Resurgent	4	5	1	5	3	9
Inflection	0	2	0	0	2	2
Total	9	9	6	6	6	18

This table describes the number of observed occurrences of each behavior type across the 18 posts selected for observation. Columns describe the traits identified in the post, and rows describe the behavior group.

Table IV.iii.b.2. A secondary breakdown of the observed behaviors, divided by complete selection criteria.

	Low view, negative sentiment	**Low view, neutral sentiment**	**Low view, positive sentiment**	**High view, negative sentiment**	**High view, neutral sentiment**	**High view, positive sentiment**	**Total**
Expected	3	1	1	2	0	0	7
Resurgent	0	2	2	1	3	1	9
Inflection	0	0	0	0	0	2	2
Total	3	3	3	3	3	3	18

IV.iii.b Post selection 1: Expected behavior

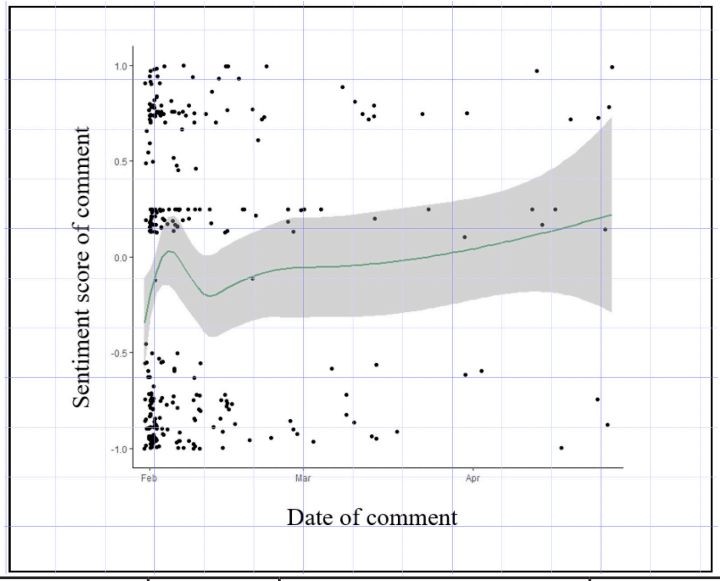
> Selection criteria: #based post 1: low view count, negative sentiment
> Content description: Meme format, movie clip. <15-second-long clip from a popular movie with white overlay text that tells the audience that the creator would do the same thing as the movie character if they were in that situation.

IV.iii.c Post selection 2: Viewership resurgence

> Selection criteria: #politics post 2: low view count, neutral sentiment
> Content description: Meme format, self-filmed video. Content creator films video of self making a reaction face. Text overlay describes an intentionally reductive political observation.

IV.iii.d Post selection 3: Sentiment inflection example 1

> Selection criteria: #PetsOfTikTok post 6: high total view count, positive sentiment
> Content description: <30-second video of a pet dog digging into the creator's furniture in a comedic way.



Total views:	1,200,000	**Average sentiment score:**	-0.565145136
Total likes:	182,600	**Like rate:**	0.152166667
Total comments:	942	**Comment rate:**	0.000785
Total shares:	1,454	**Share rate:**	0.001211667

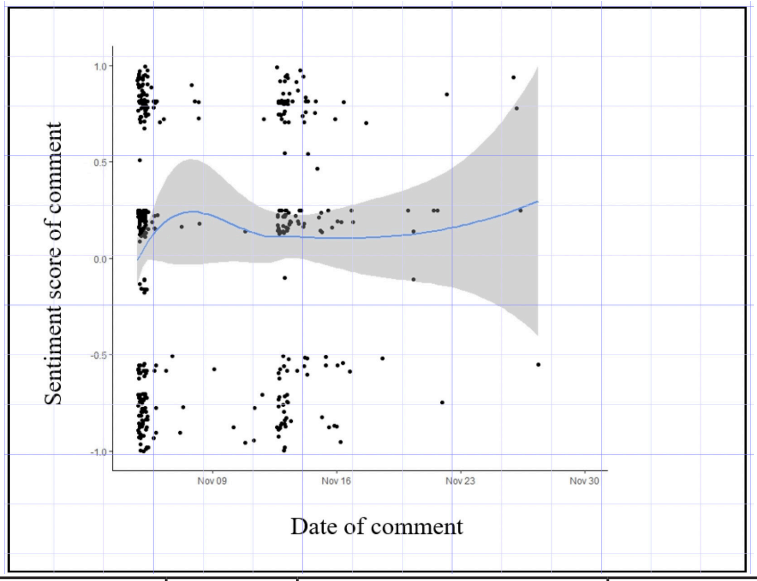
Figure IV.iii.c. The expected outcome of time series analysis, where each plotted point represents a single comment. This figure exemplifies the first behavior observed, which is considered the expected behavior. A high volume of comments can be seen clustered at the origin date. These comments are well-distributed across sentiment levels. Comment frequently falls off steeply in the months following the origin date, and sentiment distribution remains similar to the initial exposure. The exemplified front-loaded viewership and equal sentiment distribution over time align closely with expected standard recommendation system behavior.

IV.iii.e Post selection 4: Sentiment inflection example 2

> Selection criteria: #politics post 6: high view count, high sentiment
> Content description: <15-second self-filmed video of the creator gesturing to overlay text. The overlay text encourages viewers to comment on the video and share their opinions on political parties in the United States.

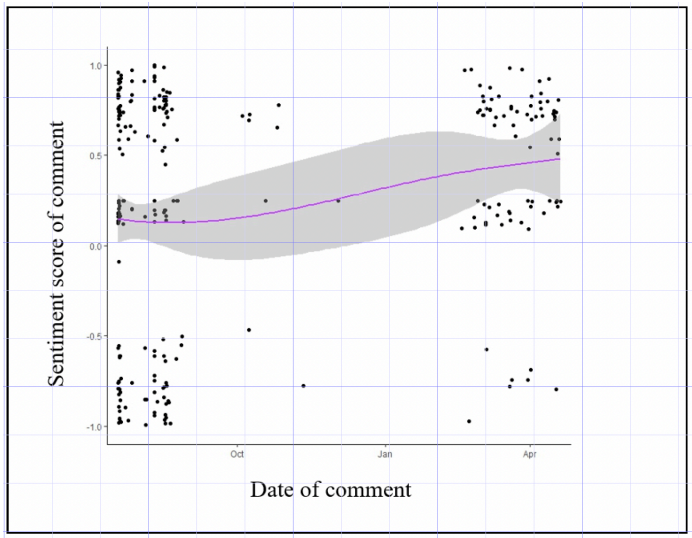
IV.iv. Recommendation algorithms and the anti-preference theory

Considering the understanding of TikTok's recommendation system that is developed in sections IV.i through IV.iii, an opportunity arises to hypothesize a connecting theory. In these sections, this paper discusses how audiences exist algorithmically, how engagement-based evaluations introduce inaccuracies, and how algorithmically-driven processes can lead to sentiment sequestering. Combining these premises, this paper suggests the theory of the *anti-preference* phenomenon as an explanation for the widespread user experience of "reaching the wrong audience." The anti-preference theory is the proposed name for a new algorithmic phenomenon that causes a creator's content to suddenly and inexplicably become saturated with hostile viewers and barraged with hostile comments. The anti-preference phenomenon occurs under the specific conditions of an algorithmic audience that shares a common content aversion and has a high propensity to comment hostilities. When these two conditions are met, the algorithm will misinterpret a high rate of hostile comments as a high success rate for the content. The content is then shown to more users within the hostile audience, where it will experience the same high comment rate. As this process repeats, the content enters a feedback loop where hostile engagement begets further hostile engagement. As the content is exposed to increasing users within the hostile audience, the original creator can become barraged with hateful and inflammatory interactions. The anti-preference is a natural deficit in engagement-based recommendation system design. Fundamentally, it is the consequence of the system's design at two important points. First is the recommendation system's inability to discern between positive and negative



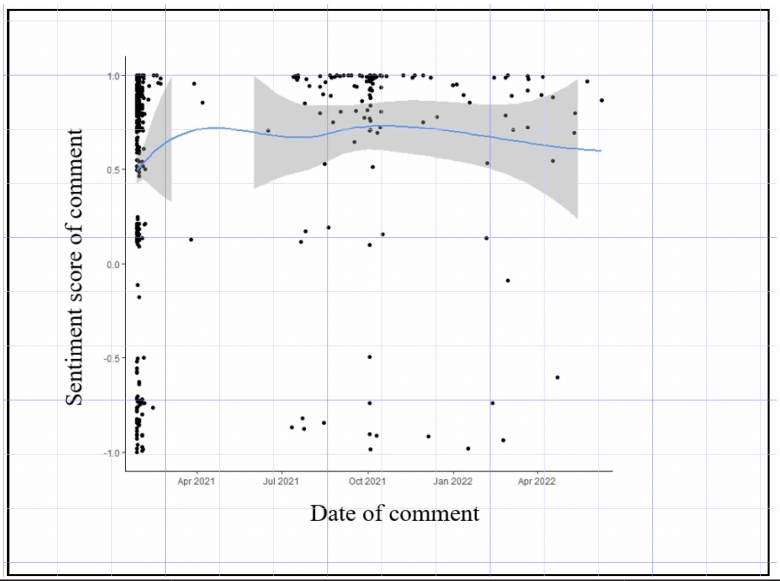
Total views:	3,400,000	**Average sentiment score:**	0.004149777
Total likes:	692,300	**Like rate:**	0.203617647
Total comments:	7,240	**Comment rate:**	0.002129412
Total shares:	7,477	**Share rate:**	0.002199118

Figure IV.iii.d. An example of algorithmic resurgence after a post-origin date is where each plotted point represents a single comment. This figure exemplifies the second behavior observed, which is considered an algorithmic resurgence. A high volume of comments can be seen near the origin date with an even distribution across sentiment. Following this, there is a section of low comment volume. Then, around ten days after the origin date, there is a second occurrence of high comment volume. The second wave of comments appears to have a similar even distribution as the first wave of comments. A total of 41 extreme outliers were removed from this figure so that the trend would be easily visible.



Total views:	52,100,000	**Average sentiment score:**	0.541364572
Total likes:	12,000,000	**Like rate:**	0.230326296
Total comments:	77,300	**Comment rate:**	0.001483685
Total shares:	303,100	**Share rate:**	0.005817658

Figure IV.iii.e. Example 1 of partial algorithmic inflection after post date, where each plotted point represents a single comment. This figure exemplifies the third behavior observed, which is considered a sentiment inflection. A high volume of comments can be seen at the origin date distributed evenly across sentiment levels. This is followed by a fall-off in comment frequency for 7–8 months. After that period, there is a month-long resurgence of comments that are sequestered to neutral and positive sentiments only. This long period of low comment volume combined with a comment resurgence that remains sequestered to a specific sentiment is consistent with the expected characteristics of an algorithmic sentiment inflection.



Total views:	5,800,000	**Average sentiment score:**	0.637500571
Total likes:	1,100,000	**Like rate:**	0.189655172
Total comments:	105,800	**Comment rate:**	0.018241379
Total shares:	35,100	**Share rate:**	0.006051724

Figure IV.iii.f. Example 2 of partial algorithmic inflection after post date, where each plotted point represents a single comment. This figure exemplifies the third behavior observed, which is considered a sentiment inflection. Like Figure 4.3e, a high volume of comments can be seen at the origin date distributed evenly across sentiment levels. This is followed by a fall-off in comment frequency for 8–9 months. After that period, there is a month-long resurgence of comments that are sequestered to positive sentiments only. This long period of low comment volume combined with a comment resurgence that remains sequestered to a specific sentiment is consistent with an algorithmic sentiment inflection.

comments. This is important when using engagement rates to gauge the success or failure of a recommendation, as it can lead to inaccuracies in a recommendation. Second is the tendency to enter feedback loops in response to high-engagement audiences. These feedback loops result in the sentiment-sequestering observed in IV.iii and the theoretical sequestering around hostile interactions that the anti-preference theory proposes. The anti-preference theory suggests that a recommendation system can build entire audiences on aversion rather than preference. In application, the anti-preference theory can be considered the hypothesized cause of the algorithmic phenomenon of "reaching the wrong audience." However, if proven true, the anti-preference theory can also have application across the design structure of all social recommendation algorithms, especially those that heavily utilize collaborative filtering and engagement-based content evaluation.

> Hypothesis: Creator-made content on TikTok can become "stuck" to hostile audiences through an algorithmic shortcoming where a viewer's negative comment is interpreted as a successful recommendation. This causes the content to be shown to an increasing number of hostile viewers with a similar high propensity to leave a negative or hateful comment. This results in a negative feedback loop where content creators are barraged with hostile comments.
> P1: TikTok's recommendation system operates in "expanding audiences."
> P2: A creator's desired or undesired content outcome can be linked to the sentiment of comments.
> P3: The undesired content outcome occurs through an algorithmic process that expands the content's audience based on a viewer's propensity to comment and has an observable effect on comment sentiment.

V. Conclusion

This paper tests a formalized version of the common algorithmic folk theory of "reaching the wrong audience." The formalized hypothesis states that "Creator-made content on TikTok can sometimes become sequestered to hostile audiences through an algorithmic failure, resulting in a distinctly undesired outcome for the creator of the video." The first premise claims that TikTok's recommendation system operates in "expanding audiences." With a cross-sectional analysis of interaction rates (IV.i), this paper partially verifies this claim. The analysis found evidence that TikTok's recommendation system continuously exposes content to new viewers who then give their preference signals, effectively "expanding the audience." However, the folk theory structures these audience expansions as groups of preference ("algorithmic audiences"), which is a claim that could not be verified. The second premise claims that A creator's desired or undesired content outcome can be linked to the sentiment of comments. With a large-scale sentiment analysis (IV.ii), this analysis verified this claim and created a framework for evaluation. The analysis found that sentiment reactions from viewers were evenly distributed across each content type analyzed. This suggests that sentiment is independent of content type and implies that content success or failure instead depends on the intent of the content creator. With this information, this analysis created a framework that evaluates the success or failure of a post in terms of sentient reactions. Using this framework, the analysis tested the third premise, which claims that this specific phenomenon of negative content outcomes occurs through an algorithmic process. After identifying 18 videos for time-series analysis, the analysis found that 11 videos displayed a significant "resurgent" behavior, where a video's viewership would become "resurrected" days, weeks, or sometimes months after its initial post date. This "unnatural" content behavior implies the involvement of some form of algorithmic process. Furthermore, 2 of the 11 videos displayed a unique form of resurgent behavior where the resurgent group expressed a homogenized sentiment that varied from the initial sentiment reactions of the same video. This was deemed a *sentiment inflection* and proved the existence of an algorithmic process that can sequester a video to a specific audience. That being said, the homogenized sentiments identified were positive groupings, which only further the success of the video (IV.ii). In contrast to the algorithmic folk theory, the analysis could not verify any harmful outcomes created by this algorithmic phenomenon. Overall, this analysis partially verifies the algorithmic folk theory of "reaching the wrong audience." Verifying this folk theory implies the existence of a new recommendation system deficiency, which this paper deems the *anti-preference theory.* The anti-preference theory proposes that when specific conditions are met, a recommendation system can build algorithmic audiences on aversion rather than preference. This theory is presented as the probable explanation for why content on TikTok "reaches the wrong audience." With the future potential of the TikTok Research API, this paper can serve as a starting point for any researchers hoping to complete more advanced statistical or mixed-method analyses on this concept.

VI. Bibliography

Abidin, Crystal. "Mapping Internet Celebrity on TikTok: Exploring Attention Economies and Visibility Labours." *Cultural Science Journal* 12, no. 1 (December 31, 2019): 77–103. https://doi.org/10.5334/csci.140.

Ashbridge, Zoe. "How the TikTok Algorithm Works: Everything You Need to Know." Search Engine Land, December 21, 2022. https://searchengineland.com/how-tiktok-algorithm-works-390229.

Chew, Shou. "TikTok's CEO on its future -- and what makes its algorithm different." TED2023, April 21, 2023. 39:23, https://www.ted.com/talks/shou_chew_tiktok_s_ceo_on_its_future_and_what_makes_its_algorithm_different/c.

Guinaudeau, Benjamin, Kevin Munger, and Fabio Votta. "Fifteen Seconds of Fame: TikTok and the Supply Side of Social Video." *Computational Communication Research* 4, no. 2 (October 1, 2022): 463–85. https://doi.org/10.5117/CCR2022.2.004.GUIN.

Jones, Corinne. "How to Train Your Algorithm: The Struggle for Public Control over Private Audience Commodities on Tiktok." *Media, Culture & Society* 45, no. 6 (September 1, 2023): 1192–1209. https://doi.org/10.1177/01634437231159555.

Karizat, Nadia, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. "Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance." *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (October 18, 2021): 305:1–305:44. https://doi.org/10.1145/3476046.

Lloyd, Andrew. "A Plus-Size TikToker Faced a Wave of Abuse after Her Travel Tips for Larger People Went Viral. But the Hate Only Pushed Her to Keep Speaking Up." *Insider*, April 6, 2023. https://www.insider.com/plus-size-tiktoker-abuse-hate-travel-tips-viral-2023-4.

Merrill, Jeremy B, and Will Oremus. "Five Points for Anger, One for a 'Like': How Facebook's Formula Fostered Rage and Misinformation." *The Washington Post*, October 26, 2021. https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/.

O'Callaghan, Derek, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. "Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems." *Social Science Computer Review* 33, no. 4 (August 1, 2015): 459–78. https://doi.org/10.1177/0894439314555329.

Rach, Markus. "A Qualitative Study on the Behavioral Impact of TikTok's Platform Mechanics on Economically Driven Content Creators." *International Journal of Social Science and Humanity*, November 4, 2021, 146–50. https://doi.org/10.18178/ijssh.2021.V11.1055.

Riemer, Kai, and Sandra Peter. "Algorithmic Audiencing: Why We Need to Rethink Free Speech on Social Media." *Journal of Information Technology* 36, no. 4 (December 1, 2021): 409–26. https://doi.org/10.1177/02683962211013358.

Smith, Ben. "How TikTok Reads Your Mind." *The New York Times*, December 6, 2021, https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html.

Thorat, Poonam B., R. M. Goudar, and Sunita Barve. "Survey on Collaborative Filtering, Content-Based Filtering and Hybrid Recommendation System." *International Journal of Computer Applications* 110, no. 4 (January 16, 2015): 31–36. https://doi.org/10.5120/19308-0760.

Thorburn, Luke. "When You Hear 'Filter Bubble,' 'Echo Chamber,' or 'Rabbit Hole'—Think 'Feedback Loop.'" *Understanding Recommenders* (blog), April 4, 2023. https://medium.com/understanding-recommenders/when-you-hear-filter-bubble-echo-chamber-or-rabbit-hole-think-feedback-loop-7d1c8733d5c.

Zeng, Jing, and D. Bondy Valdovinos Kaye. "From Content Moderation to Visibility Moderation: A Case Study of Platform Governance on TikTok." *Policy & Internet* 14, no. 1 (2022): 79–95. https://doi.org/10.1002/poi3.287.

Zhang, Min, and Yiqun Liu. "A Commentary of TikTok Recommendation Algorithms in MIT Technology Review 2021." *Fundamental Research* 1, no. 6 (November 1, 2021): 846–47. https://doi.org/10.1016/j.fmre.2021.11.015.