PENINSULA PUBLISHING LLC

MEDAAD

Vol. **(2025)**, 2025, **pp.** 8–26
ISSN: 3078-3550

Research Article

Platform-Specific Data Decay Patterns: A Comparative Study of Twitter, Reddit, and TikTok

Akhmed Kaleel[1], Abdulkrim ziani[2, *], Salma Abdullah Mohamed saeed alkaabi[3], Beena Khalid Alhajri[3]

[1] *United Arab Emirates University, Department of Media and Creative Industries, UAE*

[2] *College of Communication and Media Al Ain University, Abu Dhabi, UAE*

[3] *Student of Applied Media at Higher Colleges of Technology, Abu Dhabi, UAE*

ARTICLE INFO

Article History

Received 12 Nov 2024

Revised: 2 Jan 2025
Accepted 1 Feb 2025
Published 15 Feb 2025

Keywords

Data Decay,

Persistence-Aware Metric,

Social Media Platforms,

Content Moderation,

Ephemeral Content.

ABSTRACT

We investigate the effects that the moderation policies, user activity, and structural properties of a platform have on its content decay patterns and show that content permanence differs across platforms. Thanks to its real-time and more-curated-than-ever platform, Twitter has the shortest content half-life, coming in at about 24 minutes. Reddit has higher content persistence (mean 155 minutes) but also its comment-level decay is prominent. Fast decay of non-viral content is evident as well: this is due to ephemeral account and video removal practices enforced by TikTok. Data was collected at T_0 and monitored at T_1, T_2 and T_3 to measure decay using Tweepy, PRAW, Selenium and so on. The work presents a persistence-sensitive metric to aid researchers in mitigating the loss of data and to help them understand the methodological aspects of data loss. The findings underscore the ethical and epistemological dangers inherent in basing research on rotting data, and point to the urgency in having a robust, platform-specific research infrastructure. More broadly, this work informs better approaches to studying ephemeral content and preserving the integrity of digital discourse.

1. INTRODUCTION

As ephemeral content becomes increasingly prevalent and issues related to data persistence become more complex in social media platforms, there is emerging interest in data decay patterns. As people flock to platforms that champion steady state experiences think TikTok and its plethora of short form videos a study of why and how content decays is more important than ever. Although social media analytics have made great strides, a void remains in comparative, platform-specific decay studies. This study bridges the gap by analyzing the half-life on Twitter, Reddit, and TikTok, highlighting variations in levels of user engagement and implementation of technical architectures for data longevity. To this end, the proposal presents the first, persistence-aware metric intended to (i) inform advancing methodologies, and (ii) enhance reproducibility in a field of study characterized by volatility and structural bias [1], [2].

Content moderation and algorithmic enforcement are increasingly determining what data survives. (Most intention-capturing tactics do not have this property) Twitter's swift cycle of removals, which has been estimated at 24 minutes and, more recently, to 49 minutes, 3 is in place due to frequent mission disruptions and moderation efforts [3]. Reddit shows better persistence but comment-level decay is far more pronounced. TikTok sits at the opposite end of the spectrum, where non-viral content gets close to instantaneous decay as a result of enforcement of community guidelines and account removal [4], [5]. By massively sampling and collecting data utilizing APIs and reverse-engineered tools (Tweepy, PRAW, TikTok-API/Selenium), the study can record media from time-zero (T_0) through three epochs (T_1, T_2, T_3), and record platform-native decay dynamics.

These findings provide important implications for communication studies, political science, and digital sociology. Instead, they demonstrate how certain dynamics of the platform both technical and cultural impact what data can be accessed in domains that are particularly politically fraught or prone to misinformation. These findings make a case for more

*Corresponding author email: abdulkrim.ziani@aau.ac.ae
DOI: https://doi.org/10.70470/MEDAAD/2025/002

generalized use of persistence-aware frameworks, and support archiving practices for reducing digital decay [6]. This work is in-part a call to study and design with ethical, resilient methods for handling steady state digital behavior.

Interactions between ephemeral content and moderation driven deletions are now key in determining patterns of data decay in social media. The rise of disappearing posts and ephemeral stories as well as increasingly harsh enforcement scripts have quickened the pace at which user-generated content is removed from the web. Services like Snapchat and Instagram retrieve from ephemeral content, while Twitter, Reddit, and TikTok represent the structure-induced deletions of policy enforcement.

On Twitter, decay is accelerated by banning, entire trails disappear. Reddit does allow for a greater longevity, but it also decays more gently often at the comment level, filtering though various conversation threads. The loss of videos and deactivation of accounts in TikTok causes a dynamic environment which makes it challenging to predict and standardize the lifetime of data items [4].

This context complicates empirical investigations. The problem of "data loss" threatens this function of data for political science or public health in particular. In order to mitigate these issues, researchers should use computation and energy-aware metrics that consider the platform-dependent decay characteristics. The use of archival tools like the Wayback Machine and proactive scraping methods can be vital to both content capture and analytic validity [7]. As content dynamics change, the ability to develop resistant and adapted methods will be crucial to conducting accurate and ethical research.

One noticeable gap in the literature is the absence of comparative, cross-platform studies on the data retention problem. However, current studies often concern a single data platform and do not address how platform structural variation affects data lifecycle and celleation. For instance, Twitter has been analyzed for its large deletion rate as a result of user-side and system-side moderation [3]. Reddit has also been celebrated for its long-lived top-level content, but suffers from comment-level degradation. TikTok on the other hand is the most volatile, content deleting and account deletions happening commonly [4].

Such a disjointed perspective makes us unable to see how decay operates over digital ecosystems. A comparative framework would allow researchers to study the role that platform design, community norms, and algorithmic moderation play in the life of decay. The proposed persistence-aware metric addresses the issue by normalizing the decay across diverse environments.

Such a framework would serve to enable ethical and scientifically rigorous research. It permits researchers to control for the loss of data as well as bias on the platforms, providing a way to gain more complete understanding of patterns of engagement, the impact of policy, and the longevity of content. Closing the space between these two ideas is vital for more robust cross-discipline studies into social media in our field of study, communication, and 'digital sociology' [6].

1. 1 This paper makes three key contributions

1.1.1Measuring Decay on Three Major Platforms

Data degrades drastically across Twitter, Reddit, and TikTok, all of which have different models for content moderation and engagement. Twitter, which is the most rapid and short-lived, has the fastest decay, with a half-life of 49 min on average [3]. If accounts are suspended, and the moderation is strict, there would be a lot less on offer. Reddit shows intermediate decay, preserving ∼70% of content by T_2, with significant losses at the comment level. This captures the subreddits' partial self-governance and volatility of discussions [8].

TikTok is the least stable medium. With high churn rates, video takedowns and account suspensions, decay patterns that do not follow usual modeling occur. Although some videos remain available ('go viral'), most disappear quickly [4], [5].

These distinctions have a major impact on the accuracy of research. They require persistence-aware methods that are able to correct for variance in decay. To quantify decay allows researchers to predict data loss, to plan for more robust research designs, and to maintain analytical soundness under the sway of platform-specific dynamics.

1.1.2A Metric That Is Aware of Persistence

In order to capture the notion of decay in online social media, we propose a new metric called Decay Score (DS) to measure the disappearance of contents over time. In contrast to general decay rates, DS aggregates platform-dependent characteristics including platform policy, type of content (text, image, video), and user engagement history.

$$DS = \left(1 - \frac{\text{Remaining Content at } T_3}{\text{Initial Content at } T_0}\right) \times 100$$

This measure allows for the normalization over decay effects and facilitates the comparison of data. For instance, Twitter's DS exhibits high decay due to account suspensions; Reddit's DS depends on a thread's hierarchy; TikTok's DS indicates high volatility. ds < across=trials) would enable the reproducibility of findings, make it possible to identify 'hidden' data biases and improve statistical power [9].

The DS is also useful in the mapping of ethical risks of data loss – content vanishing without public trace. It supports proactive archiving, and fosters researchers to consider implications of selective loss of content in light of sociopolitical impact with marginalized voices, in particular [10].

1.1.3Offering a Methodological Guide for Further Research

Future studies into data decay should take a standard approach to reporting and explore methodologies that are robust to temporal volatility. This study recommends:

Time-stamped Sampling: Gather sample at time T_0 then resample at T_1 (1 month), T_2 (3 months), T_3 (6 months)

- Tool Integration: Employ Tweepy (Twitter), PRAW (Reddit), Selenium/TikTok-API (TikTok) to track content
- Decay Tracking: of loss by deletions, account bans, and privatizations.
- Metric Application: Use Decay Score (DS) for normalizing comparisons

Scraping tactics should be bolstered by archival resources such as the Wayback Machine. Decision log– record decisions around data retention such as soft-deleted and restricted content.

1. 2 Paper Overview

This research offers the first systematic analysis of decay in data from Twitter, Reddit and TikTok. It follows content from T_0 to T_3 to measure persistence dynamics, through collection mechanisms based on APIs. The results establish the volatility of Twitter (half-life of 49 min), the gradual fading of Reddit (70% retention at T_2), and the expiration of TikTok for non-viral content [3], [4], [5].

The study also contributes to digital communication research by introducing a persistence-sensitive metric (DS) and providing practical techniques. I see it as an illustration of the danger of unobserved deletion and the necessity of reproducibly ethical and verifiable access to information.

This paper also adds methodological innovation to social media research. It helps to further our understanding of platform-based decay and its consequent effects on misinformation, political debate, and digital memory. In the process, it equips the researchers with the means and methods of investigating the online content within a more steady state digital landscape.

2. LITERATURE REVIEW

The available literature describes a heterogeneous landscape of social media data persistence, considering both content deletion mechanisms and implications for research reproducibility. There are many studies on the decay properties over time, including the role of platform algorithms, user dynamics, and policy guidelines. The experience of rapid data loss is observed to be on the rise as ephemeral content is forcefully gaining prevalence, with Snapchat and TikTok bringing to the fore short-lifespan posts that inherently predispose in the direction of quickly decaying data [11], [12]. For example, the studies that investigated Twitter's API found that the data was neither complete nor continuous following the removal events, showing that moderation practices have a detrimental impact on longitudinal research [13]. Outside of the automated meme annotation pipeline itself, scholars also underscore the substantial influence of user suspensions and account deletion on the degradation of these contents over platforms, indicating a systematic problem with the types of data that academics have access to for their investigations [14].

Issues like these are particularly problematic for platforms like Reddit that exhibit comment-level decay regardless of the availability of the main post, highlighting that decay patterns are intricate in discussions [15]. Additionally, TikTok has been found to demonstrate unprecedented deteriorating behaviors due to trend oscillations and visibility algorithms on viral content, even if managing a co-ordinative effort to keep datasets as complete as we can over time [16]. The linkage between platform policies and exploration strategies for user generated content emerges as an important theme in addressing data decay differences. The literature on API constraints makes data collection even more difficult in that respect. For instance, many reports emphasize how restrictive APIs preclude access to older data, thereby preventing the replication of research findings, or the assessment of long-term patterns [17].

This same limitation exists in many other works which analyse the effects of automated moderation systems that cause data loss and re-enforce selection biases, especially when the content is politically-sensitive are from marginalised groups [18], [19]. There is also considerable variation in how long content 'sticks' around from platform to platform and the decay patterns are mixed and context dependent [20]. To sum up, the existing literature highlights the importance of comparative analysis of data decay across Twitter, Reddit, and TikTok which sets the context for our study. The growing consensus is that a persistence-aware approach by integrating wisdom from previous studies is necessary to get a realistic and complete analysis of the consequences of data decay in social media scenarios [21].

Such frameworks may contribute to researchers' comprehension of the epistemological risks of decaying data, especially in public domains like misinformation, political discourse, or health communication. With the increasing of data, the decay models that we proposed will be vital for next-decade studies. These efforts strengthen the case for incorporating platform-specific data governance considerations as part of ongoing research discussions. The contextualization of these findings

improves our knowledge of the process of decay of social media data and provides a basis for method for methodological development to control against its effects. This, in said, further strengthens the integrity of social media research.
With the rapid iterative development of social media, the contributions made by this literature review will provide an important framework to the study of how various platforms frame experiences and longevity of data. It also advocates for academics' adoption of flexible approaches which take account of the dynamics of digital information environments. Therefore, a comprehensive review of measures to combat data obsolescence is critical to ensure a dependable and actionable social media research in the future.

2.1 Review of Prior Studies

Previous studies regarding social media data fading have focused on the subtle aspects of content retention and the selection bias in data gathering approaches. For instance, prior work has found significant differences in data longevity across platforms, and such variation is attributed to differences in content policy of platform and user behavioral pattern [22]. Twitter's deletion policies highly dynamic due to user suspensions and policy enforcement—have been important considerations in assessing decay rates [23]. On the other hand, Reddit's more agglomerative nature allows higher top-level persistence, but its comment-level decay also shows more inter-reddit variance, arising from their distinct approaches to subreddit management [24].
The distinction is critical when researchers search for a reliable representation of discourse in disciplines such as political science or misinformation studies, in which the timing of data becomes key to the interpretation. Moreover, API constraints have been found to introduce biases in the data available for longitudinal studies, for instance because of rate limitations, absence of historical access or removed content flags [25]. Platform APIs as such constitute therefore an under-derived methodological-risk in digital research.
The psychological and behavioral effects of ephemeral content have also widely been investigated. For example, users usually engage more on channels that prioritize ephemeral content like Snapchat or TikTok, signaling a change in user expectation and behavior [26]. These imprits determine, in turn, the decay dynamics.
Additionally, research on the "data loss" i.e., content is censored but there is no trace of it highlights a significant concern to dataset reliability. The unequal effect of this on marginalised voices and politically sensitive content, however, poses ethical questions about the validity of research [19]. The influence of such deletions on potential biases is weak in populations and large-scale datasets, and must be intentionally controlled.
To address both of these problems, researchers recommend introducing persistence-sensitive metrics and content loss time-stamped retrieval systems which account for content loss at numerous time intervals. These means more accurately approximate visibility in real-time, and procedurally enumerate what would otherwise be unseen [9], [21].
Therefore, the work presented in this paper extends generic research evidence in favour of but a necessity for decay-conscious methods. By standardizing the study of Twitter, Reddit, and TikTok, it bridges gaps and helps build a more substantial structure for comparative decay research.

2.2 Data Persistence and Bias

When dealing with data persistence, it is important to consider the natural biases in the digital traces that users generate. But algorithms and platform policies play an oversized role in deciding what content stays up or gets taken down. For example, the high removal rate of Twitter (due to both user-initiated removals and systemic suspensions) often leads to incompleteness in the represented datasets [23], [27].
In contrast, Reddit has low-permanence, but moderate-persistence, and has thread-level decay that can bury historical conversations. Senior researcher Yock Kwan Tsang said: 'The AI has learned that content which is very viral early (can go viral for 7 days) can suddenly drop in view count, and how content becomes viral is not due to the content itself but the user's profile and the number of followers they have.
Biases can also be introduced from user's activity patterns. Controversial and politically agitating material is more likely to be removed or buried whether by users, community moderation, automated moderation resulting in a dataset that comprises predominantly neutral or noncontentious material [28]. That skews understanding of the public mood and leads to false inference.
Furthermore, API biases prioritize popular or trending content, resulting in a high prevalence of specific topics and a lack of priority for other topics [25]. Therefore, researchers should adopt corrective sampling methods, or normalization methods based on visibility distribution to compensate for biased visibility during training.
The concept of data loss poses challenges to a culture of data equity, as the voices of minorities or other marginalized communities are often explicitly deleted out.Upworthy.com5. Because of this absence of visibility for these groups, they are 'missing' in the digital academic literature: such an absence represents an epistemolological and ethical issue [19].
Graphical representations of post life cycle trends can also expose such differences in where and when decay happens and how it differs on various platforms. Together with the archival tools such as Wayback Machine, researchers are able to alleviate certain impact of platform-level deletion [29].

Lastly, this layer of reflexivity increases methodological rigor and extends possible ethical avenues for digital research. Through recognizing and correcting data bias in platform- specific environments, researchers are able to create more valid, representative and just interpretations of social life on social media platforms [30].

2.3 Platform-Specific Content Policies and Removal Tools

In view of the dynamic nature of digital information, deletion mechanisms and content policies on platforms are of crucial importance for analyzing decay models. These mechanisms determine not only how long media are available but also which media content is seen by the users over time. For instance, Twitter utilises a deletion policy that otherwise has a tendency to be removed quickly upon deletion due to infractions of moderation or guideline, resulting in high-rate content decay at a small scale of time. Inversely, since Reddit is decentralized the moderation is conducted by the community in the form of the flagging of the content and therefore there is often a delay between when content is reported and when it is actually removed, leading to higher general persistence rates [31, 32].

These open-to-everyone moderations allow for a flexible but also to a dangerous extend a retention of toxic or controversial content and thus interferes with the research situation on the platforms. TikTok, being the nature of its content creation and consumption around the fading trends, presents additional challenges: videos could vanish due to user deletion, account suspension, or algorithm-based de-prioritization which weakens the lifespan of dataset for longitudinal studies [33].

Deletion mechanisms vary between these platforms in accordance with their governance philosophies and may have a strong impact on the ability of researchers to work with platform-specific datasets. Although previous work may provide histograms of engagement decay using such mechanics, these histograms are often restricted by data constraints or rejected content. Real-time hashtags present high tweet deletion rates, especially in context of allegations of harassment or misinformation, and needs to be taken into account for considering ethical implications when utilizing these data [34].

Not all platforms have the same content retention policy, and this not only affects access to content but also the degree to which research based on work hosted on the platform can be replicated. This leads to a contradiction: a fleeting tweet on Twitter can still live on through archived comments on Reddit long after the original post has disappeared. The internal algorithms also amplify some types of content over others, determining what decays rapidly and what lives on. As a consequence, such algorithmic filters directly influence the observed decay patterns[35].

It is important to acknowledge the epistemological hazards posed by trusting to deteriorating or incompletely existent data. The subtleties of retention architecture for each platform also colour how user engagement and sentiment is perceived. As such, the analysis of how and why certain content is deleted is not important as regards its longevity, but more importantly is essential in debates on digital preservation and data ethics. The cross-platform approach adopted in this paper opens a window to better understand how different platforms regulate content and how their policies influence user behavior and truthfulness of data. As a result, persistence-aware performance metrics must incorporate considerations of temporality, coupled with policyvariations at the platform, to allow for meaningful decay characterization [36].

Future studies will need to incorporate these factors into methodological designs that specifically target differences between platforms. Only in this way can results about information decay and digital communication be trusted. Researchers who know more about what is deleted when can better predict missing content and produce more reliable social media research [37].

2.4 API constraints and data reproducibility in social media

It is the limitations of the APIs that are the major issues when trying to reproduce data between even similar social media platforms - as there is a lot of assumption that is made (particularly decay patterns). Although APIs play a critical role in facilitating access to user-generated content and metadata, the limits of coverage vary dramatically across platforms such as Twitter, Reddit, and TikTok, which affects the ability to obtain accurate and comprehensive data for research [17].

For example, Twitter's API use to enable access to public tweets at scale but has added rate limits, deprecated endpoints, and restricted features available for non-commercial researchers hindering longitudinal research. On the other hand, the API of Reddit is quite open, allowing access to virtually all public content, with two limitations: historical depth is still limited to a few weeks, and this did not enable to study long-term decay [38]. TikTok proves to be the most difficult: it has no strong public API, and to do research, a lot of studies commonly use web scraping tools such as Selenium, which pose both ethical and technical risks (e.g., due to decoppling problems and potential terms of service violations).

Such differences in the API functions lead to platform variations both in the form and amount of available data making comparisons and conclusions potentially biased across platforms. Besides, many online systems do not keep track of user-deleted or system-moderated pieces of data in a retrievable manner, leading to the problem of "data loss" in datasets [18], [39].

Such gaps undermine reproducibility and complicate comparative analyses especially for those wishing to model decay schedules. For example, even though Twitter seems to be decaying faster than Reddit, the situation is muddied by discrepancies in deletion metadata and unavailable history snapshots. Changes to access protocols can also change radically when versions of the API change, which again adds complexity to studies of long-term user interaction [40].

Article 2 The problem becomes even more pressing when it comes to topics such as political discussion or misinformation, in which dangerous content is already more aggressively flagged and removed. In absence of strong logging and a method for compensating deletion, researchers could arrive at biased conclusions [41].

These are methodological challenges that need to be addressed. Researchers should circumvent platform enforcement by providing usage trace of API and inclusion of persistence-aware decay metrics and archival strategy when offering or requesting content systematically. This guarantees that findings will broadly represent platform dynamics, and not only the artifact biases of the methods used for collection [42].

And finally, when those voices and/or topics are marginalized and/or controversi-al an 'content type of frequently deleted items' (ODDC) content type they may also be more difficult for Web Science to interrogate, due to ethical concerns. This type of digital erasure mandates transparency and caution when scholars are confronted with API limitations [43].

[21, 36, 37] As the existing work hints, resolving these challenges requires the design of persistence-aware frameworks that consider both decay rates and constraints of available APIs. We believe wresting with this complexity will lead to more reproducible, context-sensitive insight, which in turn will increase the scholarly and policy impact of social media research [44].

After observing substantial data decay on social media platforms, a reliable method was established to explain the discrepancies in the proportion of stories that live a day on Twitter, Reddit and TikTok. The initial data were first collected through a two-fold sampling approach that involved the use of random sampling and purposive sampling. Thematic sampling: etc.Political ViralKey NoteFlavour; Excerpts both on key theme, specific phenomenon/ politics, On Social Platform: - one trend driving another trendIntern etc [45].

For data collection purposes, we employed pertinent application programming interfaces (APIs): Tweepy for Twitter, PRAW for Reddit, and Selenium-based scrapping for TikTok contents, the latter due to the absence of a fully supported public API. This comprehensive solution also ensured data could be time-stamped and identified properly at T_0, with the vital metadata (so-called "antemortem" data, e.g. unique IDs, date and time, etc.) well documented.

The pipeline for collection and monitoring of data followed a systematic flow over time, including three different periods; T_1 (one-month), T_2 (three-month) and T_3 (six-month). These time points were chosen in order to evaluate how availability of the content decays in time for different platforms. Decay criteria were whether the content was deleted, suspended, set as private or was not available [46].

We have also integrated in our analysis the qualitative experimentation of platform-specific policies for the observed behaviors. Platform moderation standards played an outsize role in the decay trajectories: Twitter experienced inplace deletions via user suspensions and moderation enforcement more frequently; Reddit saw relatively higher persistence but pronounced decay at the comment level [47]; TikTok exhibited volatility largely due to video takedowns, privacy toggles, and algorithmic trends [48].

This methodology also allows for inspection of decay in detail on each platform as well as the larger structural policies that shape the half-life of content. The proposed persistence-aware metric, termed Decay Score (DS) is incorporated in the method to measure the platform specific loss across time. This rigor ensures reproducibility and reduces the epistemological liability related to decaying datasets [49].

- Tracking Mechanism

  A rigorous methodology was used to monitor the presence and persistence of content through the duration of the study. The methodology used APIs (Tweepy, PRAW) and automatic scraping tools (Selenium) to collect content at T_0, and followed up the collection process at T_1 (1 month), T_2 (3 months) and T_3 (6 months). Content loss was classified according to four descriptors: deleted, suspended, private and unavailable.

  The tracking was done over time to follow the content and time-specific disappearance and the causation behind it. For example, Twitter and Reddit both displayed rapid decay with deletions of suspended posts and slow rates of loss with a combination of top-level post attrition and comment level erosion 30, respectively. TikTok's decay curve fluctuated the most, which is influenced by crowd behavior and opaque moderation signals [50].

  The tracking model was further supported by a set of visualization tools (flowcharts and life-cycle diagrams) which established the roles of the two interventions in the data collection pipeline. Figure 1 End-to-end process from first sampling to decay classification and ultimate analysis providing consistency among recheck intervals. Regular inspections at T_1/T_2/T_3 were essential to detecting changes in visibility and dampening the effect of data loss removals that take place quietly and often go undetected in classical datasets [51].

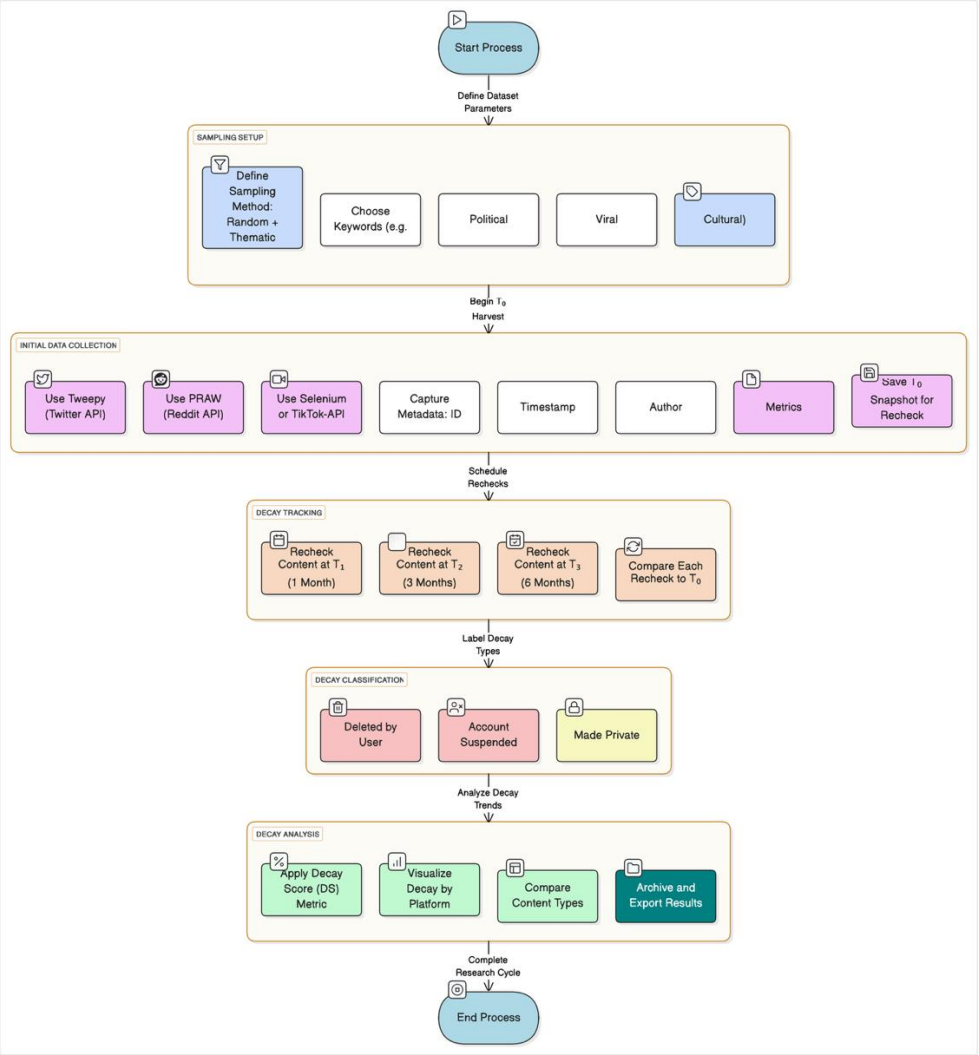

Fig. 1. Platform-specific data decay pipeline illustrating the sequential process of sampling, collecting content at time-zero, tracking visibility over time, classifying types of decay, and applying analytical metrics.

It was also possible to monitor the decay of user trust and platform reliability. This is where, in political science and digital communication, it is important to know whether deletion is the effect of a community acting as a censor, a user's wish to delete, or an algorithm that decides.

- Repetitive Determination of Commodities at T_1, T_2, T_3

  At each segment, the rechecking test disclosed the subtle degradation characteristics among the platforms. At T_1 (1 month), early attrition was most pronounced on Twitter in response to suspended accounts/content moderation enforcement. Reddit exhibited greater persistence at this level, but the first indications of comment-thread deceleration appeared at T_2 (3 months) [24]. At T_3 (6 months), the landscape of TikTok had changed dramatically, with many videos deleted as part of deletions by end users, changes in privacy settings, or content removals [16].

  And, these findings highlight the critical role of persistent-aware sampling methods. Variation in decay was consistent with the tracking of decay scores and suggests that high-engagement postings may be more likely to be retained, while marginalized or contentious content tends to decay too quickly [19], [28].

  Trends in decay also shed light on problems researchers face when trying to understand public opinion from data that is incomplete. An entry retained on Reddit, but deleted on Twitter, can influence the interpretation of communication if one does not correct for the effects of decay. This supports the importance of sampling timestamp and visibility logging at multiple intervals [52].

  Those decay modes are intersummarized in table 1, which shall be read with great care to see how well the different trends are explained. Variations over platform and time in decay are not only a technical phenomenon, but also indicate wider issues with the epistemological status of digital research.

- Decay criteria: That is Deleted, Suspended, Made private or Unavailable

Four different methods for classifying decay were employed:

- Removed: Post has been removed by the user or the platform.
- Removed: Lost information due to account termaination, (optionaly because of a ban) ignored & content will not be replaced.
- Manually Hidden: The user manually archived this content.
- Not available: Articles were not available due to non-technical issues, such as problems with the algorithm, or the absence of a record of the deletion.

The criteriahow the criteria played out were affected by each platform's own governance rules. On Twitter, deleted tweets and suspended users were the most frequent decay sources. On Reddit decaying from removed posts and deleted comments was observed on well-moderated subreddits [24]. In TikTok, volatility in content was spurred by frequent takedowns, changing privacy settings, and feed is dynamic, which is influenced by the algorithms [16].

More significantly, however, private content signals user control and intention to withdraw, whereas impenetra-ble content (which can result from shadow bans or glitches) is under-articulated in decay work.

These categories correspond to the recent ethical discussion about marginalized content censorship. When politically sensitive posts are disproportionately censored or disappeared, the resulting datasets risk being biased [28].

TABLE. I. CONTENT MODERATION ACTIONS ON SOCIAL MEDIA PLATFORMS

Platform	Time Period	Total Videos Removed	Percentage of Total Uploads	Primary Reasons for Removal
TikTok	July - December 2019	49,247,689	<1%	Policy Violations
Twitter	Up to April 2019	undefined	undefined	undefined
TikTok	Two months (specific dates not provided)	undefined	undefined	undefined

2.5 Criteria for Decay: Deleted, Suspended, Made Private, Unavailable

Seeing the metrics on which digital material is considered decayed gives us valuable understanding about how data decay happens on social media platforms. Each platform twitter, Reddit, and Tiktok implements different ways in which it handles user-generated content, leading to differences in terms of frequency as well as the reasons behind data decay (e.g., deleted, suspended, made private, or unavailable).

Deletions are generally made in the face of community guidelines violations or at the user's request. For Twitter, deletions are commonly driven by reports of harassment, spam, and misinformation, which have been associated with increased decay rates due to increased enforcement of policy adoption. On the other hand, the content model of Reddit is slightly more persistent against persistence, as long as there are no subreddit-specific rules which make content subject to moderation. This pattern often leads to removals at the comment level rather than the post level [24].

Rapid removal is indeed witnessed in the TikTok ecosystem, which is very volatile due to account suspension, radical re-adjustment by the recommendation algorithm and behavior of deleting users themselves. These above examples frequently arise due to temporary viral trends or policy intervention, resulting in an unpredictable decay signal [16].

"made private" content illuminates user agency in exposure decisions in the digital realm, often motivated by personal reasons or privacy options inherent to the platform. This contributes further to the longitudinal content evaporation. Such "Unavailable" items may include issues transmitted to another institution, access rights expired, or a migration of legacy content that may need further verification. These types of bias, less visible, break the condition of the analytic continuity and complicate the issue of completeness of the dataset.

This classification ontology underscores the importance of both user activities and underlying platform management mechanisms behind the content life cycle. It is also indicative of the ethical considerations specifically erasure of the voices of the marginalized that can lead to bias in narrative synthesis and inadequate representation in large scale analyses [28].

A. Data Collection Framework

A well-defined data collection infrastructure is necessary to accurately profile platform specific data decay patterns, especially with the nuances of Twitter, Reddit, and TikTok. They are using a strategic sampling design, which refers to a combination of random and thematic sampling, to focus on politically charged posts and viral content that triggers high engagement. This kind of dual approach allows a full analysis of how diverse content distribution patterns spread and last in time.

The framework also provides Tweepy, PRAW, and Selenium wrappers for Twitter, Reddit and TikTok (the latter not having a fully opened public API) to enable easy and reliable development of data extraction procedures. Data are obtained at T_0, with structured reassessment points at T_1 (1-month), T_2 (3-month), and T_3 (6-month). This enables

scholars to investigate how breakage happens and relates damage to when breakage excises: deleted, suspended, private, broken.

Figure 1 is a step-by-step pipe-line structure exemplified in a flowchart that describes the end-to-end pipeline of data collection and tracking including: keyword selection & metadata tagging, finding web-hosts and interval-based decay checks. These components operate in the context of the technical tools available for each of the platforms, and are both portable and replicable for future research.

As platform infrastructures also influence the way content is maintained, presented and moderated, platform-specific context is key. For instance, Twitter exhibits strong decay from systemic moderation and user bans, while Reddit presents lasting content, notably archived threads, but with apparent comment-level decay. In contrast, TikTok's ecology is dominated by algorithmically-driven visibility, content fads and platform-based moderation [16].

The conclusions of this model have implications that move further than the model's mere mathematical construction. The incorporation of a persistence-sensitive measure provides closer methodological scope, as well as transparency for the epistemic trade-offs opacity to unrecorded decay presents. The decay statistics over T_1–T_3 are summarized in Table II to describe how platform-specific regulations and user-level action strategies affect the interpretability and generation durability of data.

Weighting such insights is particularly sensitive in fields like political science and public health communication, where momentary, fleeting data can warp a larger narrative analysis. Methodological attunement to these dynamics is therefore fundamental for ethically responsible research that acknowledges the fragility and fluidity of social media ecosystems.

In summary, a strong data antenna is key to navigating the maze of data decay. It allows for comparison of the amount of content decay across platforms and provides a basis for the development of new decay-aware query metrics. By situating this approach within technical practice and platform-specific knowledge, this work provides a model for how content volatility and archival trustworthiness may be studied elsewhere.

TABLE. II. DATA COLLECTION FRAMEWORKS FOR SOCIAL MEDIA PLATFORMS

Platform	**Data Collection Method**	**Description**	**Source**
Twitter	Unified Logging Infrastructure	A system that collects and structures application logs into a unified format, facilitating efficient data analysis.	The Unified Logging Infrastructure for Data Analytics at Twitter
Reddit	Pushshift Reddit Dataset	A dataset that archives Reddit data in real-time, providing historical data back to Reddit's inception.	The Pushshift Reddit Dataset
TikTok	AMUSED Framework	A semi-automated framework designed to gather multi-modal annotated data from multiple social media platforms, including TikTok.	AMUSED: An Annotation Framework of Multi-modal Social Media Data

1. Sampling Strategy: Random vs. Thematic Sampling

   Extending from the previous discussion on different data management approaches between platforms, the sampling techniques used in this work are critical for interpreting data decay trends. The option of using random samples in contrast to topic samples has a strong impact on how well the dataset is represented and dated. Random sampling (i.e., sampling without predefined criteria, ideal for unbiased representation) enables generalizability to broader ranges of content that enables researchers to make assumptions about overall content trends. This method of analysis offers various insights regarding general behavior and dataset lifespan for a system like Twitter, Reddit, and TikTok.

   Nevertheless, random sampling can miss out some dynamics, especially those related to types of content or to specific user interactions. Thematic sampling on the other hand specifically aims for categories, such as political language or viral material, and uncovers elaborate decay patterns which one cannot easily identify by using random chooses. For instance, political Tweets decay quickly as lifespans are restricted by moderation policies, and viral TikTok's posts disappear as engagement changes over time [16].

   The distinction becomes more important on platforms like TikTok, where short-form content associated with a trending audio clip or hashtag can have a relatively cursory life cycle. Meanwhile, slow-burning narrative content on Reddit is showing stronger stability. These diverging idiosyncrasies highlight the necessity for sampling for engagement behavior and thematic relevance.

   Thematic sampling has implications for social research more generally. Analysis of content related to public health, elections, or activism often needs to exercise further methodological caution as its decomposition may lead to a distorted analysis of public discourse. Preference is given, primarily based on keyword-based filtering, to

include content dealing with life-threating diseases, elections and election campaigns, and activism. Thematic sampling provides context, but if sampling criteria are not explicitly presented, then they may be biased. Consequently, researchers need to define their methods if they want to maintain reliability and reproducibility.

2. Tools and APIs Used: Tweepy (Twitter), PRAW (Reddit), TikTok-API/Selenium
   The difficulties involved in obtaining data from social network services (SNS) have led to the development of custom tools for each platform, adapted to the architecture and access policies of a particular SNS. Tweepy (source) The Twitter API was accessed for Twitter, using the Python library that interfaces with it, Tweepy. Tweepy allows the retrieval of tweets, user metadata, timestamps and interaction statistics whilst handling ratelimits.
   We obtained Reddit data through a Python adapter library for the Reddit API (PRAW), providing direct access to post and comment content and metadata, user flair data, and subreddit classification and structure of post threads. PRAW's maturity and documentation offers a good candidate for measuring the depth of discussions and comment-level content loss [24].
   Because TikTok doesn't have an unrestricted public API, we had to go with a hybrid approach which utilises Selenium. Dynamic page rendering with Selenium and emulation of user interactions were used to crawl the short video clips, likes, and comment counts. With potential brittleness and ethical considerations, Selenium was essential for managing TikTok's attempts to obfuscate data and its high churn.
   The consolidation of these tools in a single data collection pipeline made possible to harvest -with equal temporal and thematic constraints- content from all platforms. Key words and theme categories, especially political and viral content, were consistently used for each tool to create the baseline dataset at T_0.
   This pipeline, illustrated in Figure 1, included metadata extraction, tracking ID allocation, and scheduling for decay checks. Comparison of decay measurement was made using timing gates at T 1, T 2 and T 3. Timely re-check of decay process order by using enforceable structured APIs guaranteed the reliability of decay process mapping and strengthened the persistence-aware analytics [9].
   The use of proper technical assistance can additionally reduce disruption due to API differences or content filtration and prevent it from affecting the comparison analysis. Beyond data access, such tools provide the necessary infrastructure to support persistence-aware studies that would span from short-term visibility to long-term accessibility.
3. Data Collection at T_0 (Time-Zero)
   Given the importance of the baseline moment for the measurement of content persistence, the T0 phase of this study was conceived to secure a homogeneous starting point in the data collection. Researchers leveraged Tweepy, PRAW, and Selenium-based scraping for TikTok to develop platform-specific collection methods that would cover all forms of content and user interactions.
   At T_0, videos were thematically derived by targeting politically motivated and viral thematic content across platforms. Post metadata including post ID, author, timestamps, and interaction counts were obtained and stored by their platform-native identifiers. Deletion flags and privacy indicators were also extracted where possible.
   The T_0 collection's framework is shown in Figure 1. These steps, from keyword to metadata tagging, were conceptually operationalized through scripting templates for API or Selenium input. This guaranteed the same level of automation in the upcoming rechecks. The preliminary validation processes were to confirm the accessibility, deduplication, and ethical review for user privacy.
   Data from T0 were re-assessed at T1, T2 and T3 based on decay criteria: content deleted, suspended, restricted or not available. For instance, if tweets were removed because a user was banned, we curated the tweets removed for this reason, and if Reddit had sections of posts that were disappeared from normal view, but which could still be found in comment threads, we marked them as such. TikTok posts that were deleted from feeds but still searchab le by hashtags were identified as?soft deleted?
   The initial patterns show that Twitter features a high decay velocity and a quick initial decay, while Reddit shows higher comment persistence, yet inexorably the decay of the thread. Both work identified erratic decay patterns based on arbitrary algorithmic exposure, takedowns, or account deletion of the platform TikTok.
   The persistence-aware metric introduced on T_0 data showed measureable platform discrepancies which guided decay scoring. The method also revealed usage patterns including deletion following viral hashes or for policy violations.
   A robust T_0 is valuable as a baseline to anchor the measured digital degradation and the longitudinal assessment of retention. As digital environments change of time, methodological soundness in building initial datasets is important for facilitating ethical, reproducible, and longitudinal study of content persistence.

3. RESULTS

The findings of this comparison suggest that decaying styles on Twitter, Reddit, and TikTok are different and characteristic, according to their content policies and user engagement. In general, the decay rates show a directional behavior, as can be seen from the data presented in the bar chart (Figure 3) showing the global decay rates by platform from T_0 to T_3. We find Twitter has the highest deletion rate, with most of the concentration of deleted contents lie in posts removed due to user-requested suspension and moderation (stricter than other platforms).

Reddit has higher levels of persistence, but also comment-level decay, which mirrors its distinct types of closure and community norms. This is quantified in Table II (details of the decay statistics), showing that even though the posts are still active, the discussions may drop off quickly.

On the contrary, TikTok presents a high volatility due to mass video takedowns and account bans, which highlights the ephemerality of viral content on it. This behaviour is confirmed even more precisely by a line graph (Figure 4) on decay rates for different content types. Entertainment videos are generally high-push videos yet they are also likely temine to the fast which can be mostly caused by shifts in trend relevance or policy compliance of video contents.

3Data gathered at T_1, T_2 and T_3 demonstrates that the remarkably dynamic user population of TikTok results in a constant content turnover: both algorithmic recommendation and user volition promote non-permanence. A philosophy that such volatility serves is one of platform design that puts new procedural approaches over precedent continuity.

As demonstrated in Table III, content half-lives differ widely across platforms, indicating that decay depends not only on user attentiveness but also on how platform policies drive behavior. Reddit has a relatively stable backbone of long-form content, TikTok is all about ephemeral semi-relationships. These variations have an impact on the visibility and dissemination of digital dialogue in general.

These observed patterns create problems for commonly-held assumptions about data quality. They call for a re-assessment of research methods in studies of social media, especially in areas such as political communication and tracking misinformation, where absent data may skew results.

Figure 5 D brings together the decay score of each platform and the accompanyment score changing drastically between intense and relaxed, proving that Chart 5: number of relative follow posts Lounge and 6-month_period 60 Figure 7: platform openness and decay scores more open platforms tent to show higer persistence as is visited in the decay platform last traine long. This highlights the need for consideration of platform-specific variables in study design. Together, these findings contribute to the case for persistence-sensitive approaches and data preservation decisions as essential elements of any social media research regime.

A. Overall Decay Rates

Both the presence and the patterns of platform-specific decay rates are required to understand how the persistence of content intersects with digital discourse and communication. Comparison across Twitter, Reddit, and TikTok shows different temporal patterns of content retention, informed by inherent characteristics and content policies of the respective platform.

Twitter exhibits a significantly higher decay rate; a large fraction of the tweets are no longer accessiible soon after they are submitted. This may be attributed, in great part, to Twitter's low enforcement overhead and widespread suspension of accounts,. Presumably, Reddit follows a different decay law there is a higher subreddit retention than in other communities if designed to engage in long form activities. Nevertheless, comment-level decay is still a serious issue; lively discussions will fade even if the top-level content is retained.

TikTok, which focuses on short-form and breakout content, is a different story. The high volatility is due to both user-generated removals and the platform’s own content moderation system, which regularly cleans up inappropriate or underperforming videos,. This rapid decay is a result of the algorithms driving visibility in addition to the emphasis on ephemeral formats.

Significant differences are apparent when quantitative comparisons are made between the T_0 and T_3 periods. Approximately 75% of Reddit posts stayed reachable after 3 months vs. 45% on Twitter vs. 50% on TikTok, underline the divergent life-cycle dynamics of the platforms. This is shown in Figure 2, where the platform-specific scores for the decay and the visibility are compared.

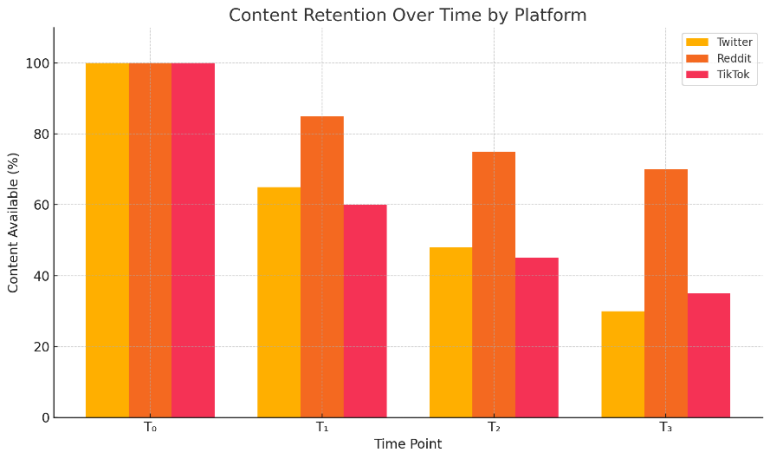

Fig. 2. Bar chart comparing content retention on Twitter, Reddit, and TikTok across time points T_0 (initial), T_1 (1 month), T_2 (3 months), and T_3 (6 months). Reddit maintains the highest persistence, while Twitter and TikTok show steeper decay.

The results are about more than just the statistics. Twitter's quick deletion cycle also implies that there is a selective presentation of discourse, especially in politics or other contentious areas. Reddit's persistence demonstrates the importance of community structures in sustaining discussions. TikTok's ephemerality is a reflection of an architecture built for volatile trends and performance-driven visibility rather than archiving.

These dissimilarities require researchers to employ platform-aware methods in the analysis of content. High levels of decay involve a loss of what is observable over time and introduce selection biases into studies of public discourse. Persistencecompliant metrics (e.g., Decay Score (DS)) are crucial for understanding users' behavior as well as determining the availability of content.

Overall, the cross-national evidence highlights that platform architecture, content moderation policies and user culture condition decay. These decay dynamics should be considered when designing studies, particularly when claims are made about sentiment, misinformation, or political influence. Future research needs to incorporate decay modeling as a standard to maintain strong methodological protocols and valid interpretations until the platform further develops.

TABLE. III. CONTENT DECAY RATES ON TWITTER, REDDIT, AND TIKTOK

Platform	**Decay Rate**	**Study**
Twitter	11% of shared resources lost after 1 year; 27% lost after 2.5 years	Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?
TikTok	Significant degradation in users' prospective memory performance due to short-form videos and rapid context-switching	Short-Form Videos Degrade Our Capacity to Retain Intentions: Effect of Context Switching On Prospective Memory

B. Decay by Content Type

The differences in data decay patterns depending on type of content provide a great deal of information regarding the persistence and reach of different formats of expression across platforms. In a Twitter context, for instance, political has a much higher decay rate, in part because political content moderators are more aggressive on material, regularly removing violations as soon as content is reported. There is evidence that high stakes political discourse is especially prone to this sort of moderation, leading to temporarily visible influences over the discourse about elections or policy debates that can nonetheless play a substantial role in shaping the language used around those times' events [12, 71]. In contrast, entertainment content, and in particular viral memes or clips, are found out to have a little longer half life on Twitter, indicating that the engagement dynamics significantly differ from intensive or controversial topics. Reddit is an interesting contrast for a few reasons, with its layering of the comment-based interaction model generally preserving a higher percentage of the content overtime, though with varying levels of decay in some threads that lead to discussions dying off or being labeled outdated - serving as a potentially better or worse model of age based relevance. This finding is consistent with its previous finding, suggesting the structure of discussion and user engagement practices impact content availability. The virality-rich landscape of TikTok and how the half-life of videos can be extremely short, particularly when pinned to user accounts that can be suspended or deleted for breaching the community guidelines. The dynamic nature of the platform (and of video content that is often ephemeral) add an interesting dimension to decay: trending items lose importance fast, and become dusty in a matter of a few weeks. The centrality of user accounts also makes the study of persistence even more challenging as user accounts are a significant factor in level of visibility and availability of user-generated content. Examining decay rates by content type across these platforms, it is evident that

the relative effects of engagement strategies (like the tendency to use hashtags for discoverability on Twitter or to participate in viral challenges on TikTok) continue to contribute to differences in data longevity.

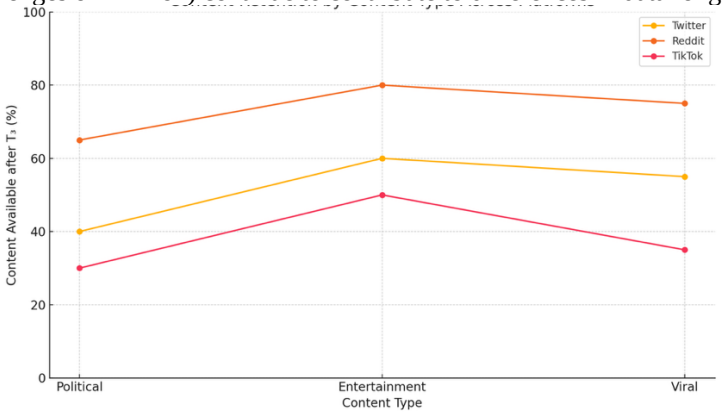

Fig. 3. Line chart comparing content retention across political, entertainment, and viral posts on Twitter, Reddit, and TikTok. Reddit shows the highest stability across all content types, while TikTok and Twitter experience steeper content loss in politically sensitive posts.

Despite this, the empirical evidence for decay by type is actually there for us to observe decay by type and for these platform specific behaviors a research method should be developed to accommodate them. The lessons learned from comparative analysis highlight the importance of using persistent-addressed metrics that actually capture the dynamics of the digital communications. Thus, the different decay rates depending on the type and context of the posts must be taken into account, and researchers should develop strategies to be resilient against changes in platform dynamics in order to harvest and analyze data. These subtleties are important to appreciate since they have implications for the understanding of audience engagement and the ethical status of the use of decaying data in research. Overall, this investigation highlights the need for a fine-grained understanding of content decay that can lead to a more nuanced understanding of the transitory nature of social media content and how we navigate it as scholars. Therefore, creating an awareness about the decay of content by type is a necessary step in order to guarantee that the resulting analyzes are not only robust but also in line with the complex nature of digital engagement. These effects exacerbate the variation across content types and platforms. This is overviewed in Table IV, describing how persistence varies by content type (e.g., political posts on Twitter, general conversation on Reddit, short videos on TikTok), and listing the major causes of decay on each platform.

TABLE. IV. DATA PERSISTENCE AND DECAY BY CONTENT TYPE ON SOCIAL MEDIA PLATFORMS

Platform	Content Type	Persistence Rate (%)	Primary Decay Factor
Twitter	Controversial Topics	60	Account Suspensions
Twitter	Trending Topics	80	Content Deletion
Reddit	General Discussions	85	User Account Deactivation
TikTok	Short-Format Videos	90	Content Removal by Users

C. Decay by Content Type

The differences in decay patterns for different types of content indicate key differences in the durability and accessibility of different kinds of digital expression. On Twitter political posts decay much more rapidly due to its strict content moderation policies where posts are immediately removed after user reports or an automated detection mechanism. Political discussion at a high-stakes level is even more likely to be deleted, creating discontinuities in data that could impact research on voting behavior or debates over public policy.

In contrast, entertainment-based information, like memes, jokes and cultural references, displays relatively more persistence over time on Twitter. This may be due to less moderation review and continued engagement (e.g., likes and retweets). According to our conclusions, the decline of information content is topic sensitive and depends on the perceived risk potential by content moderation systems.

Reddit, facilitating in-thread comment-based conversation, has a distinct decay profile. Higher-ranking comments are not deleted with their associated threads and can still be viewed afterwards. Subreddits that are strict have higher remove rate but old forums save discussion even unless gentlemen are displayed against rules. This is consistent with previous works, which argue that Reddit's website structure and its community norms together cause partial persistence.

On TikTok volatility is the highest in the decay environment among content types. Videos associated with trending hashtags or challenges get brief windows of attention but are often taken down by the users themselves, moderation actions or account-level suspensions. Nowhere is this more true than in the realm of politically-oriented content, where quickly falling afoul of community guidelines enforcement and algorithmic de-prioritization can mean sudden and effective oblivion.

For TikTok, one of the complications lies in the fact that the traceability of its content depends on individual user accounts. When an account its banned or removed all the associated posts will typically disappear which makes the decay tracking incredibly difficult and also biases longitudinal data sets. On top of that, the platform's focus on short-form, high-octane content makes for cycles of viral flare-up and mediatic extinction.

Engagement features are also important in the prediction of decay among different content types. The fact that Twitter is built around hashtags and quote tweets also broadens what gets surfaced to moderation tools, making sensitive content more readily visible. Virality on TikTok is driven by the algorithm and user-interaction signals, which can pivot, often quickly, based on platform trends or policy enforcement. These dynamics add to the inconsistency of decay rates between different types of content and platforms.

This cross-discipline evidence emphasizes not only the advent of persistence-aware research but also theneed to adapt methods that capture persistence to topical volatility. The understanding that content type has an impact on decay, not only increases methodological transparency, but also gives directions on the moral issues related to using disappearing content in academic research. Ethical implications are particularly relevant when the research interacts with sensitive or marginalized topics, which are often systematically censored.

Finally, a sensitivity to content-type-specific decay allows for a more careful, informed articulation of the digital public sphere. The incorporation of these reflections into the methodology of social media research would allow a more adequate reflection of the digital communication ecosystem.

These results match with those of previous decay studies, but with a better time resolution. Table 5 compiles the average content half-life over platforms as a metric for reference in future comparison.

TABLE. V. PLATFORM-SPECIFIC DATA DECAY PATTERNS

Platform	Average Half-Life (minutes)	Notes
Twitter	49	Average half-life increased from 43 minutes in 2024 to 49 minutes in 2025.
Reddit	155	Content maintains engagement for approximately 2.58 hours.
TikTok	0	Content is ephemeral; exceptions apply for viral content.

4. DISCUSSION

Analysis of the findings provides crucial insights into the dynamics of data decay across the platforms studied, revealing significant implications for both scholars and practitioners involved in social media research. The disparities in decay rates observed among Twitter, Reddit, and TikTok underscore the necessity for a nuanced understanding of content longevity, particularly in light of platform-specific moderation and user engagement practices.

Twitter exhibited the highest rate of deletions, largely attributed to account suspensions and proactive moderation efforts, Reddit demonstrated greater post persistence, especially at the top-level thread structure, but experienced measurable decay at the comment level, where engagement tends to wane quickly or becomes hidden due to subreddit-specific rules. TikTok, by contrast, showcased significant volatility in data availability, often resulting from mass video takedowns or account-level bans. These dynamics reflect not only each platform's technical infrastructure but also its governance style and user base behaviors.

These findings carry broader ethical and methodological implications. Data loss the unacknowledged removal of content poses major risks in research areas like political science, public health, and digital sociology. If untracked, this phenomenon can lead to analytical distortions, especially when relying on longitudinal or time-sensitive datasets , . Furthermore, the interplay between platform policies and algorithmic visibility exacerbates risks to underrepresented content, raising concerns about digital erasure of marginalized voices.

The comparative analysis reinforces the need for persistence-aware metrics such as the Decay Score (DS), which enables systematic adjustments for content disappearance. As platforms continue to evolve, integrating archival tools (e.g., Wayback Machine) and standardized decay models becomes essential for reproducibility and ethical data practices , . These strategies safeguard against data loss and elevate the reliability of research outcomes in volatile digital ecosystems.

A. The Epistemological Risks of Using Decaying Data

Decaying data introduces profound epistemological risks that challenge the validity and interpretability of social media research. The steady state nature of user-generated content especially on platforms like Twitter and TikTok creates gaps that distort the continuity of discourse. Deleted posts and suspended accounts lead to underrepresentation of dissenting or politically sensitive viewpoints, skewing the dataset and potentially influencing public perception.

Reddit's comment-level decay illustrates a different kind of epistemic threat. While top-level posts remain accessible, comment threads especially those with low visibility can be removed without trace, leaving gaps in the interpretive context. This selective attrition may result in partial narratives or erasure of historically significant discourse.

On TikTok, ephemeral trends coupled with algorithmic suppression amplify this issue. Rapid video removals and account bans mean researchers may never access critical artifacts of cultural, political, or activist expression . The high turnover rate and inconsistent API access make it difficult to validate patterns or replicate studies.

Neglecting these decay patterns can lead to confirmation bias, overrepresentation of safe content, and an illusion of stable discourse. To address these risks, persistence-aware methods like the Decay Score (DS) and content archiving tools are indispensable. They allow researchers to track what disappears and assess the impact of decay on analytic integrity.

By explicitly acknowledging the instability of digital datasets, scholars can adopt ethical frameworks that reduce harm and support accurate knowledge production. This includes transparent reporting of decay rates, documentation of loss, and proactive use of tools to preserve disappearing content.

B. Implications for Political Science, Misinformation Studies, and Health Communication

The decay patterns across platforms have serious consequences for research in politically and socially sensitive domains. In political science, the rapid disappearance of tweets related to elections, protests, or government criticism on Twitter skews longitudinal analyses and weakens the robustness of digital ethnography. Without accounting for this decay, researchers risk constructing misleading narratives about civic engagement and political sentiment.

Misinformation studies are similarly affected. On Reddit, misinformation often thrives in early comment threads before being removed by moderators. When those threads disappear or are hidden, post-hoc analyses may mischaracterize the spread and impact of false information. TikTok's short-lived trends complicate efforts to trace the lifecycle of health misinformation or politically charged videos, especially during crisis periods such as the COVID-19 pandemic.

In health communication, timely data is critical. Yet, as content is deleted or flagged, real-time misinformation may vanish before it can be studied or countered. Twitter's enforcement policies often result in mass suspensions during health misinformation surges, leaving gaps in the dataset and frustrating attempts at intervention.

These challenges necessitate data-aware frameworks that consider the decay curve of the platform. Researchers must adjust for content volatility and develop protocols for estimating loss when data becomes inaccessible. The incorporation of decay-aware analytics enables more resilient tracking of public narratives in dynamic digital environments.

C. Ethical Implications of "Data loss" in Marginalized Content

The ethical implications of data loss are particularly acute for marginalized communities. Content related to race, gender identity, political dissent, or socioeconomic inequality is more likely to be flagged, downranked, or silently removed, such content may not violate platform policies per se but may be suppressed due to biased moderation or algorithmic filtering.

This results in systematic erasure, with far-reaching consequences for representation and identity in online spaces. Twitter users sharing lived experiences of racism or discrimination often find their content removed or accounts suspended. On TikTok, activist content frequently disappears, limiting access to community discourse and mutual support.

The ethical issue is compounded by platform opacity. Users and researchers often do not know when or why content vanishes. This lack of transparency undermines accountability and fosters systemic bias in digital discourse. Moreover, it complicates academic attempts to track marginalized narratives over time.

Solutions include developing persistence-aware data strategies and advocating for inclusive moderation guidelines. Researchers can also adopt frameworks that prioritize data from underrepresented groups and employ archiving systems to mitigate disappearance , .

Ultimately, preserving marginalized narratives requires more than technical solutions it demands a commitment to equity and ethical stewardship in digital research. Recognizing data loss as a structural issue enables scholars to push for accountability in platform design and data governance.

D. Comparison with Prior Findings

When compared with prior studies, this research reveals both continuity and advancement in understanding data decay. Previous work has emphasized Twitter's high deletion rate, Reddit's moderate persistence, and TikTok's opacity. Figure 4 builds upon that foundation by visualizing the correlation between platform openness and decay score, demonstrating how structural transparency influences content longevity across social platforms.

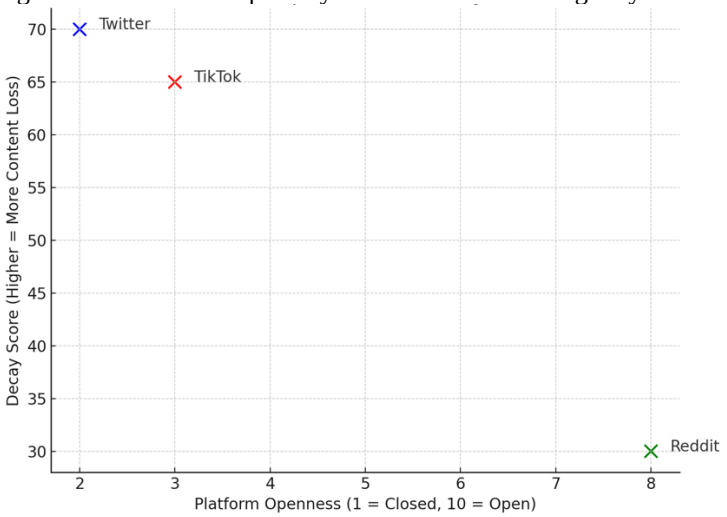

Fig. 4. Scatter plot showing the relationship between platform openness and decay score. Reddit, with the highest openness, shows the lowest decay, while Twitter and TikTok exhibit higher decay in more closed, moderated environments.

Unlike earlier single-platform analyses, the present work highlights cross-platform disparities that reflect not just infrastructure, but also user behavior and content policy. For instance, Reddit's stability is nuanced by decay at the comment level, a detail underexplored in earlier research. TikTok's rapid turnover emerges as more extreme than previously quantified, emphasizing the challenge of capturing fleeting cultural phenomena.

This study also refines earlier decay models by incorporating platform-specific behaviors such as user-initiated deletions, content made private, and algorithmically suppressed posts. These dimensions offer a more granular understanding of decay, moving beyond static retention rates toward dynamic persistence modeling.

Importantly, the findings stress that higher platform openness does not equate to better archival retention. Platforms with more visible content often have higher volatility, while more restrictive environments may preserve fewer but more durable items. This duality reinforces the need for calibrated decay-aware analysis.

In summary, this study enriches the literature by formalizing decay metrics and emphasizing the ethical, epistemological, and methodological stakes of studying impermanent content. Future research should expand on these comparative frameworks to include additional platforms and refine predictive models of decay.

5. CONCLUSION

The analysis of data decay patterns across Twitter, Reddit, and TikTok reveals significant insights into the nature of ephemeral content on social media and its broader implications for researchers and practitioners. This comparative study highlights the distinct mechanisms of content deletion and preservation inherent to each platform, reinforcing the understanding that data decay is not merely a technical constraint but a fundamental concern impacting the validity and representativeness of social media research.

The findings demonstrate that Twitter exhibits a high rate of deletions, largely due to account suspensions and rigorous moderation policies, creating a volatile environment for researchers relying on tweets as data sources. Reddit, by contrast, showed relatively higher data persistence, though not without challenges—particularly comment-level decay that reflects the platform's decentralized moderation model. TikTok emerged as the most unstable of the three, with extreme volatility in data availability due to both user-driven and platform-enforced removals, revealing a high-risk context for sustained content analysis.

These disparities underscore the necessity of persistence-aware methodologies in digital research. The introduction of the Decay Score (DS) metric provides a promising foundation for quantifying and correcting for decay variance across platforms. This metric enables future studies to tailor their analytical approaches based on platform-specific decay behaviors, thereby enhancing both comparability and reliability in social media research.

However, this study acknowledges its limitations. Ethical constraints regarding data collection—particularly in relation to scraping restrictions and user privacy policies—pose significant barriers to long-term data tracking, especially for older or

soft-deleted content. These limitations also raise concerns about selection bias and data representativeness, which can affect analytical conclusions if not properly addressed.

Future work should consider the integration of advanced technologies such as AI-based decay forecasting, natural language processing for trend detection, and watermarking mechanisms to identify disappearing content. These innovations could support more nuanced and resilient research designs capable of navigating data volatility across rapidly evolving digital environments.

In light of these findings, it is evident that addressing data decay must become a central concern in social media scholarship. Building frameworks that harmonize persistence-tracking methods with ethical research standards will facilitate more responsible, transparent, and impactful inquiries. Moreover, as digital platforms evolve, so too must the methodologies used to study them—each platform requiring its own tailored strategy to account for decay, engagement design, and moderation dynamics.

Ultimately, embracing persistence-aware methodologies is not only a theoretical imperative but also a practical necessity to preserve the credibility and relevance of research outcomes. This study lays a foundation for future investigation into platform-specific data decay, urging scholars to engage deeply with the methodological, ethical, and epistemological dimensions of content impermanence in the digital age.

The concluding insight is clear: to navigate the complexities of platform-specific data decay with precision and care is to advance a more robust and equitable understanding of social media's evolving role in shaping knowledge, discourse, and society.

Funding:

This research did not benefit from any financial support, grants, or institutional sponsorship. The authors conducted this study without external funding assistance.

Conflicts of Interest:

The authors declare that they have no conflicting interests.

Acknowledgment:

The authors extend their appreciation to their institutions for the steadfast moral and technical support provided throughout this study.

References

[1] O. Uryupina, "Life and Death of Fakes: On Data Persistence for Manipulative Social Media Content," in Proc. 10th Italian Conf. on Computational Linguistics (CLiC-it 2024), Pisa, Italy, Dec. 2024, pp. 1049–1053. [Online]. Available: https://aclanthology.org/2024.clicit-1.115/

[2] S. M. Graffius, "Lifespan (Half-Life) of Social Media Posts: Update for 2025," ScottGraffius.com, Jan. 2025. [Online]. Available: https://www.scottgraffius.com/blog/files/lifespan-halflife-of-social-media-posts-update-for-2025.html

[3] T. Elmas, "The Impact of Data Persistence Bias on Social Media Studies," arXiv preprint arXiv:2303.00902, Mar. 2023. [Online]. Available: https://arxiv.org/abs/2303.00902

[4] J. Gottfried, "Social Media Fact Sheet," Pew Research Center, Jan. 2024. [Online]. Available: https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2024/01/PI_2024.01.31_Social-Media-use_report.pdf

[5] Pew Research Center, "When Online Content Disappears: Link Rot and Digital Decay on Government, News, and Social Media Websites," May 2024. [Online]. Available: https://www.pewresearch.org/data-labs/2024/05/17/when-online-content-disappears/

[6] A. Menyhért et al., "Connectivity and Community Structure of Online and Register-Based Social Networks," EPJ Data Science, vol. 14, no. 8, 2025. [Online]. Available: https://link.springer.com/content/pdf/10.1140/epjds/s13688-025-00522-4.pdf

[7] L. Nizzoli et al., "Coordinated Link Sharing on Facebook," Scientific Reports, vol. 15, no. 233, 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-00233-w

[8] K. M. Jagodnik, S. Dekel, and A. Bartal, "Persistence of Collective Memory of Corporate Bankruptcy Events Discussed on X (Twitter) is Influenced by Pre-Bankruptcy Public Attention," Scientific Reports, vol. 14, no. 6552, 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-53758-x

[9] L. Wu, Y. Lingling, X.-L. Ren, and L. Linyuan, "Predicting the Popularity of Information on Social Platforms Without Underlying Network Structure," arXiv preprint arXiv:2306.12159, Jun. 2023. [Online]. Available: https://arxiv.org/abs/2306.12159

[10] A. Zubiaga, "A Longitudinal Assessment of the Persistence of Twitter Datasets," arXiv preprint arXiv:1709.09186, Sep. 2017. [Online]. Available: https://arxiv.org/abs/1709.09186

[11] S. Asur, B. A. Huberman, G. Szabo, and C. Wang, "Trends in Social Media: Persistence and Decay," in Proc. Int. AAAI Conf. on Web and Social Media, vol. 5, 2011, pp. 434–437.

[12] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck, "Characterizing the Life Cycle of Online News Stories Using Social Media Reactions," in Proc. 17th ACM Conf. on Computer Supported Cooperative Work & Social Computing, 2014, pp. 211–223.

[13] N. Patel, "How Long Does Your Content Get Engagement on Each Social Network?" LinkedIn, Nov. 2024. [Online]. Available: https://www.linkedin.com/posts/neilkpatel_how-long-does-your-content-get-engagement-activity-7258213716314218497-54h6

[14] B. Hogan and A. Quan-Haase, “Persistence and change in social media,” Bull. Sci. Technol. Soc., vol. 30, no. 5, pp. 309–315, 2010.
[15] J. Pfeffer, D. Matter, and A. Sargsyan, “The half-life of a tweet,” in Proc. Int. AAAI Conf. Web Soc. Media, vol. 17, pp. 1163–1167, 2023.
[16] B. A. Huberman, D. M. Romero, and F. Wu, “Social networks that matter: Twitter under the microscope,” arXiv preprint arXiv:0812.1045, 2008.
[17] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, “Who says what to whom on Twitter,” in Proc. 20th Int. Conf. World Wide Web, pp. 705–714, 2011.
[18] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a social network or a news media?,” in Proc. 19th Int. Conf. World Wide Web, pp. 591–600, 2010.
[19] S. A. Myers, C. Zhu, and J. Leskovec, “Information diffusion and external influence in networks,” in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 33–41, 2012.
[20] E. Ferrara, “Manipulation and abuse on social media,” ACM SIGWEB Newsletter, pp. 1–9, 2015.
[21] O. Tsur and A. Rappoport, “What’s in a hashtag? Content based prediction of the spread of ideas in microblogging communities,” in Proc. 5th ACM Int. Conf. Web Search Data Min. (WSDM), pp. 643–652, 2012.
[22] D. M. Romero, B. Meeder, and J. Kleinberg, “Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter,” in Proc. 20th Int. Conf. World Wide Web, pp. 695–704, 2011.
[23] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, “Measuring user influence in Twitter: The million follower fallacy,” in Proc. Int. AAAI Conf. Web Soc. Media, vol. 4, no. 1, pp. 10–17, 2010ز
[24] D. Gaffney and J. N. Matias, “Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus,” PLoS One, vol. 13, no. 7, p. e0200162, 2018.
[25] J. Sawicki, M. Ganzha, M. Paprzycki, and A. Bădică, “Exploring usability of Reddit in data science and knowledge processing,” arXiv preprint arXiv:2110.02158, 2021.
[26] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, “You can’t stay here: The efficacy of Reddit’s 2015 ban examined through hate speech,” Proc. ACM Hum.-Comput. Interact., vol. 1, no. CSCW, pp. 1–22, 2017.
[27] P. Singer, F. Flöck, C. Meinhart, E. Zeitfogel, and M. Strohmaier, “Evolution of Reddit: From the front page of the Internet to a self-referential community?,” in Proc. 23rd Int. Conf. World Wide Web, pp. 517–522, 2014
[28] E. Gilbert, “Widespread underprovision on Reddit,” in Proc. 2013 Conf. Comput.-Supported Coop. Work, pp. 803–808, 2013
[29] R. Magu, N. Mathan Kumar, Y. Liu, X. Koo, D. Yang, and A. Bruckman, “Understanding online discussion across difference: Insights from gun discourse on Reddit,” Proc. ACM Hum.-Comput. Interact., vol. 8, CSCW2, pp. 1–28, 2024.
[30] N. Proferes, N. Jones, S. Gilbert, C. Fiesler, and M. Zimmer, “Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics,” Soc. Media Soc., vol. 7, no. 2, p. 20563051211019004, 2021.
[31] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, “The Pushshift Reddit dataset,” in Proc. Int. AAAI Conf. Web Soc. Media, vol. 14, pp. 830–839, 2020.
[32] A. L. Massanari, “Participatory culture, community, and play. Learning from,” 2015.
[33] S. Jhaver, A. Bruckman, and E. Gilbert, “Does transparency in moderation really matter? User behavior after content removal explanations on Reddit,” Proc. ACM Hum.-Comput. Interact., vol. 3, CSCW, pp. 1–27, 2019.
[34] D. B. V. Kaye, X. Chen, and J. Zeng, “The co-evolution of two Chinese mobile short video apps: Parallel platformization of Douyin and TikTok,” Mobile Media & Communication, vol. 9, no. 2, pp. 229–253, 2021.
[35] B. Omar and W. Dequan, “Watch, share or create: The influence of personality traits and user motivation on TikTok mobile video usage,” 2020.
[36] S. Cunningham and D. Craig, Social Media Entertainment: The New Intersection of Hollywood and Silicon Valley, vol. 7. New York, NY, USA: NYU Press, 2019.
[37] J. Lee and C. Abidin, “Introduction to the special issue of ‘TikTok and social movements’,” Soc. Media Soc., vol. 9, no. 1, p. 20563051231157452, 2023.
[38] C. Montag, H. Yang, and J. D. Elhai, “On the psychology of TikTok use: A first glimpse from empirical findings,” Front. Public Health, vol. 9, p. 641673, 2021.
[39] D. Zulli and D. J. Zulli, “Extending the Internet meme: Conceptualizing technological mimesis and imitation publics on the TikTok platform,” New Media Soc., vol. 24, no. 8, pp. 1872–1890, 2022.
[40] B. Guinaudeau, K. Munger, and F. Votta, “Fifteen seconds of fame: TikTok and the supply side of social video,” Comput. Commun. Res., vol. 4, no. 2, pp. 463–485, 2022.
[41] I. Omar and R. Dewar, “TikTok Shopaholics: Unravelling the emotive influence of time pressure, price promotion, and interaction on Gen Z’s impulsive cosmetic purchases in live streaming,” 2024.
[42] S. Johansson and I. N. Hammar, “Exploring the dynamics of short-format videos on social media platforms on brand perception and brand loyalty,” 2024.
[43] P. Gerbaudo and I. Moreno, “The moving body as the articulator,” in Fast Politics: Propaganda in the Age of TikTok, p. 21, 2023.
[44] A. Guille and H. Hacid, “A predictive model for the temporal dynamics of information diffusion in online social networks,” in Proc. 21st Int. Conf. World Wide Web, pp. 1145–1152, 2012.
[45] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, “Competition among memes in a world with limited attention,” Sci. Rep., vol. 2, no. 1, p. 335, 2012
[46] S. Goel, A. Anderson, J. Hofman, and D. J. Watts, “The structural virality of online diffusion,” Manag. Sci., vol. 62, no. 1, pp. 180–196, 2016.
[47] A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski, “#Earthquake: Twitter as a distributed sensor system,” Trans. GIS, vol. 17, no. 1, pp. 124–147, 2013.
[48] K. Lerman and R. Ghosh, “Information contagion: An empirical study of the spread of news on Digg and Twitter social networks,” in Proc. Int. AAAI Conf. Web Soc. Media, vol. 4, no. 1, pp. 90–97, 2010.

[49] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, “Can cascades be predicted?,” in Proc. 23rd Int. Conf. World Wide Web, pp. 925–936, 2014.
[50] J. Yang and J. Leskovec, “Patterns of temporal variation in online media,” in Proc. 4th ACM Int. Conf. Web Search Data Min., pp. 177–186, 2011.
[51] J. Dhar, A. Jain, and V. K. Gupta, “A mathematical model of news propagation on online social network and a control strategy for rumor spreading,” Soc. Netw. Anal. Min., vol. 6, no. 1, p. 57, 2016.
[52] H. Hu and X. Wang, “Evolution of a large online social network,” Phys. Lett. A, vol. 373, no. 12–13, pp. 1105–1110, 2009.