

2023

DEMYSTIFYING PORTRAYALS

of Harmful Content
on YouTube and
TikTok



Weiyu Zhang
*Director, The Civic Tech Lab
National University of Singapore*

Ho Wei Yang
*Assistant Insights Manager
Truescope Singapore*

Table of Contents

CONTENTS

P01.

INTRODUCTION

P02.

OBJECTIVE

P03.

METHODOLOGY AND MEASURES

P06.

FINDINGS

P22.

LIMITATIONS AND IMPLICATIONS

P23.

RECOMMENDATIONS

FINDINGS

- *Categories of Harmful Content* p06
- *Demographics* p07
- *Hashtags Generation Techniques* p08
- *Content Curation Techniques* p09

YouTube

- *Animation* p09
- *Music* p10
- *Diaries and Vlogs* p12
- *Video Cuts/Voiceovers of TV Shows* p12
- *Sketches* p13
- *Coded Language* p14
- *Non-Human Characters* p14
- *Game References* p15

TikTok

- *Coded Language* p15
- *Music* p16
- *Diaries and Vlogs* p18
- *Video Cuts/Voiceovers of TV Shows* p19
- *Celebrity Images/Videos* p19
- *Special Characters and Symbol* p21

INTRODUCTION

Social media platforms have become the primary means for individuals to communicate, obtain information, and share content (Stasi, 2019). They afford fast, convenient access to information and the freedom to express oneself across diverse formats almost instantaneously (Ruckenstein & Turunen, 2020; Wyrwoll, 2014).

Unsurprisingly, the same platform for open expression can be used to facilitate the dissemination and amplification of harmful content. Various manifestations of harmful content exist, ranging from disinformation to hate speech and cyberbullying (Gongane, Munot & Anuse, 2022). As social media becomes an integral part of everyday lives, it is imperative to contemplate the implications it has on the well-being of vicarious consumers of social media content. Of particular concern is the younger population of social media users who are generally more impressionable during the identity development phase (Columbia University Mailman School of Public Health, 2021).

Major social media platforms have affirmed their commitment to regulating user-generated content based on their set of content moderation policies, many of which commit to safeguarding freedom of expression while maintaining a safe space for users (Facebook Community Standards; Instagram Community Guidelines; Snapchat Content Guidelines; TikTok Community Guidelines; Twitter Rules; YouTube Community Guidelines).

Notwithstanding, we continue to see harmful content circulating on social media platforms. Indeed, scholars have documented the presence of harmful content on popular social media platforms including Facebook, Instagram and Twitter (Arendt, Scherr & Romer, 2019; Dyson et al., 2016; Miguel et al., 2017; Shanahan, Brennan & House, 2019; Stänicke, 2022). More recently, two non-profit organisations have flagged a slew of harmful content circulating on TikTok - which has seen tremendous growth in its user base in recent times - and cautioning its potential implications on vulnerable populations such as children. Moreover, the reports underscore a particular concern - the role of the platform's algorithmic mechanisms which exacerbate the spread and impact of the content (Center for Countering Digital Hate, 2022; Ekō, 2023).

OBJECTIVE

Despite the commitment of social media platforms to moderate content, harmful content remains prevalent on these platforms. This necessarily implies that users have devised ways to escape platform policing. This led us to our principal objective for this report - to [demystify the creative techniques social media users adopt when posting harmful content to escape platform content moderation](#).

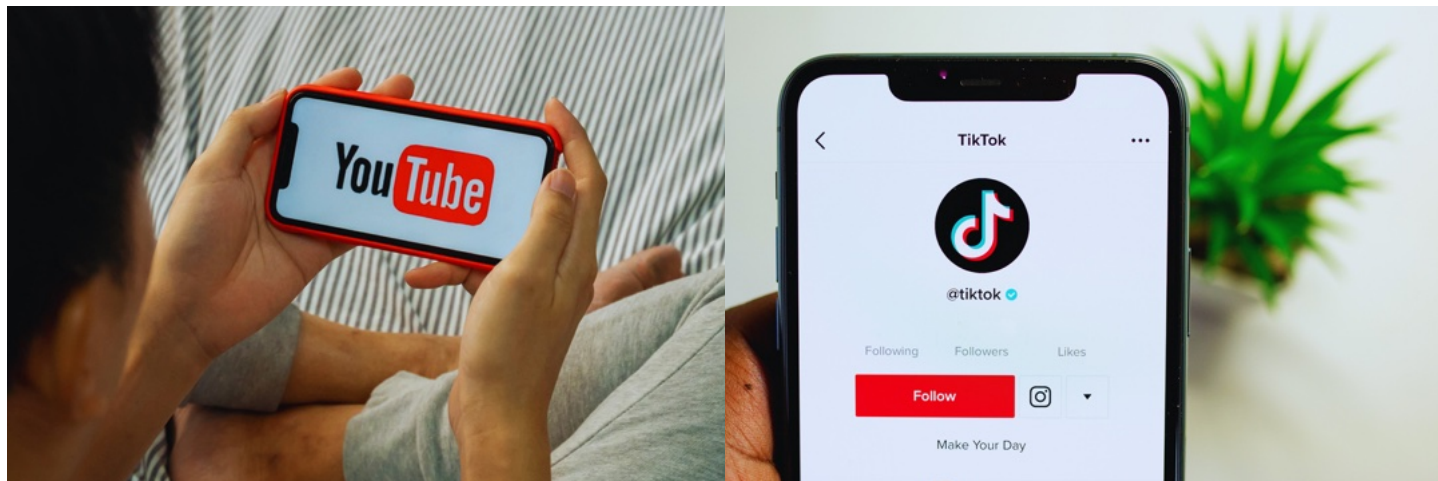
Next, we were primarily concerned with the impact of harmful content on the young, vulnerable populations. This led us to consider what forms of social media content and which platforms were more popular amongst this demographic. As short-form content popularity surges among the young population, TikTok usage continues to see unprecedented growth.

Moreover, YouTube's popularity has also seen steady growth, consistent with a longer-term shift towards video-based content (Perez, [2023](#), Pew Research Centre, [2022](#); Potrel, [2022](#); Statistia, [2023a](#), [2023b](#)). In Singapore, YouTube has a potential audience of 5.08 million (Datareportal, [2023](#)) with about 90% of users aged between 16 – 34 (Ng, [2021](#)), while there are over 1.8 million active TikTok users in Singapore of which 61.7% are aged 18 – 24 and 35.2% are aged 25 – 34 (Black Dot Research, [2023](#), Start.io, [2022](#)). With this understanding, YouTube and TikTok were selected as the primary platforms for investigation.



Based on Digital 2023: Singapore, YouTube has a potential audience of 5.08 million. In Singapore, approximately younger audiences (aged 24 and below) on social media comprise of 18.3% of all users.

METHODOLOGY AND MEASURES



There are various types of harmful content on social media, ranging from violence, hate speech, sexual exploitation to privacy breach, incivility, and scams. Because of our concern of young users, we chose to focus on harmful content that is prevalent among this user group, namely, [self-harm](#), [eating disorders](#) and [suicide](#).

Through reviewing existing academic works on the topics (Alhassan, et al., [2021](#); Wang, et al., [2017](#)), we generated an initial keyword list of harmful content for further search. Using YouTube and TikTok accounts that were established for work and research purposes, we made an initial keyword search through the platforms' search function for harmful content which may be circulating on the platforms.

The purpose of the search was to locate more commonly recurring keywords and hashtags that users may be using to categorise and label harmful content. The search yielded a list of 14 keywords and 19 hashtags ([see Table 1](#)) which were then ingested into Truescope's proprietary media intelligence platform and database for data extraction. Only English social media data was extracted from YouTube and TikTok, and the monitoring period was from [9 March 2023 to 9 April 2023](#).

A total of 610 YouTube and 508 TikTok posts were crawled and assessed for relevance for the scope of this report. 10% of all videos (61 for YouTube, 51 for TikTok) were randomly selected and coded independently by two coders to test inter-coder reliability. The inter-coder reliability (percentage agreement) stands at [0.95 for YouTube](#) and [0.91 for TikTok](#).

We first consulted YouTube and TikTok’s community guides on their definitions of “harmful content”, particularly with regards to the topics of “self-harm, suicides, and eating disorder”. Our measurements are in line with platforms’ definitions of allowable content by excluding recovery, awareness and educational videos that are deemed “hopeful” and “inspirational” instead of “harmful”.

We also excluded videos which do not explicitly showcase the users’ own intentions to engage in any form of harmful behaviour. The step of exclusion has led to an exclusion of 373 videos on YouTube and 314 videos on TikTok from analysis. The final sample of social media data consisted of 237 YouTube and 194 TikTok videos.

Table 1

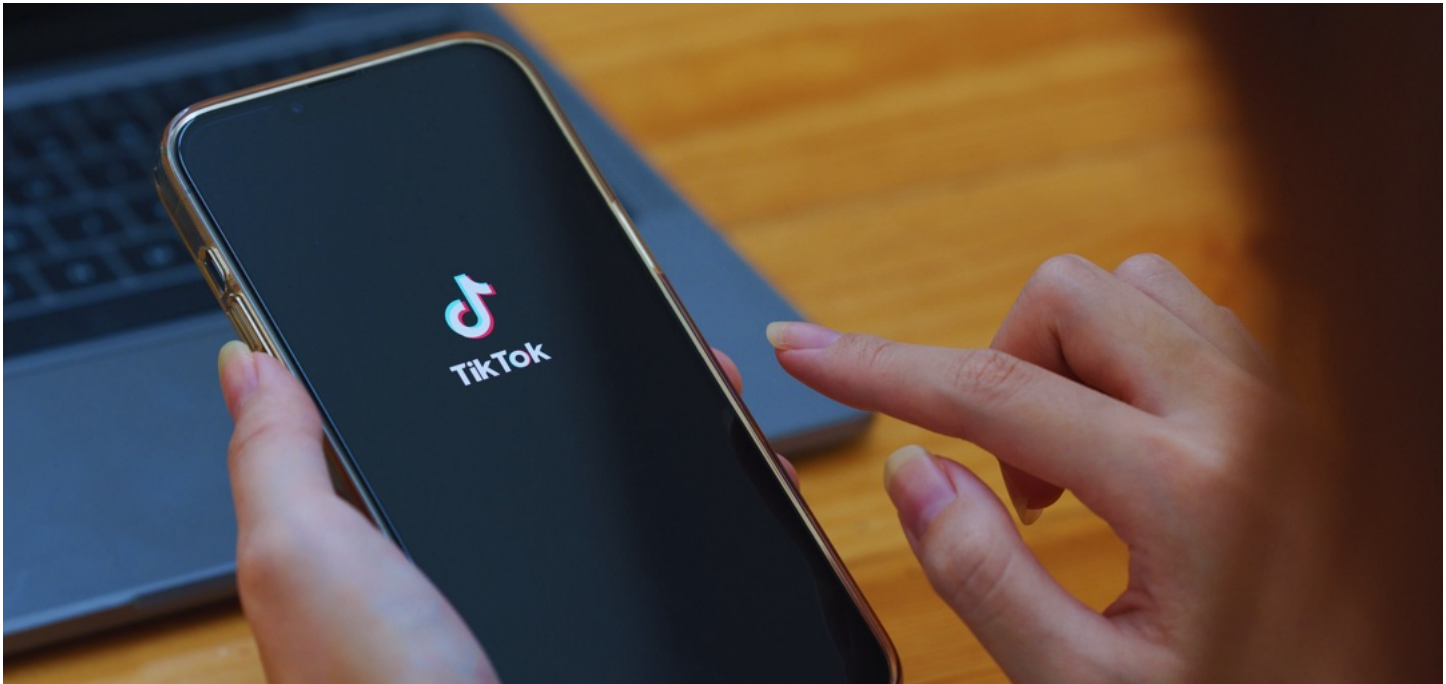
List of Keywords and Hashtags

Keywords

“junkorexic”, “junkorexia”, “thinspo”,
“tw ed”, “tw sh”, “sewer slide”,
“suwerslide”, “depreshun”,
“depression”, “self-harm”, “self harm”,
“self-injury”, “self injury”, “self cutting”

Hashtags

#bodyshame | #bodyshamed |
#junkorexic | #thinspo | #ana | #sh | #ed |
#twed | #twsh | #twvent | #shvent |
#ednotsheeran | #EdSheeranDisorder |
#imdone | #ihatemylife | #ihatemyself |
#suwerslide | #depreshun | #enditall



The final sample of social media data contained information such as the username, date of video posting, volume of engagements (i.e., reactions, comments and shares), number of views, number of subscribers/followers.

Each video was then further coded for the following attributes:

- Type of harmful content
- Presentation techniques adopted by users; and
- Demographic information

The final attribute on demographic information was done by scraping all available information in the public domain. This includes viewing the users' public profile page for self-disclosures and/or scrutinising the users' videos, wherever possible.

For videos which showed the users' body parts (e.g., hair, face, torso, hands, and/or feet), we made the best-possible attempt to identify the users' gender and age group, and avoid misrepresentations.

The final set of harmful content mostly aren't always actively promoting to consumers to mimic such harmful behaviours, but rather, they do showcase potential ways subtly—in which consumers can engage in those harmful behaviours (and perhaps, provoke those who already have an intention). Furthermore, they certainly convey the emotional pain that the users are experiencing which may be distressing for those watching.

Observational Note

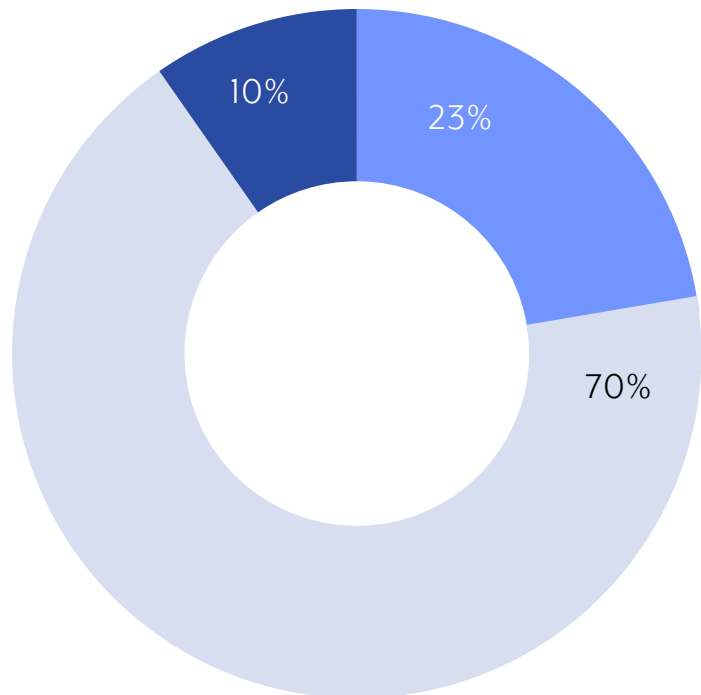
When the research team revisited our final sample of YouTube and TikTok videos to review our coding work (one month after the initial round of coding was complete), we noticed that a significant number of videos have been removed – approximately 50% on YouTube and 30% on TikTok. These may be attributed to the active content moderation by the respective platforms.

FINDINGS

Categories of Harmful Content

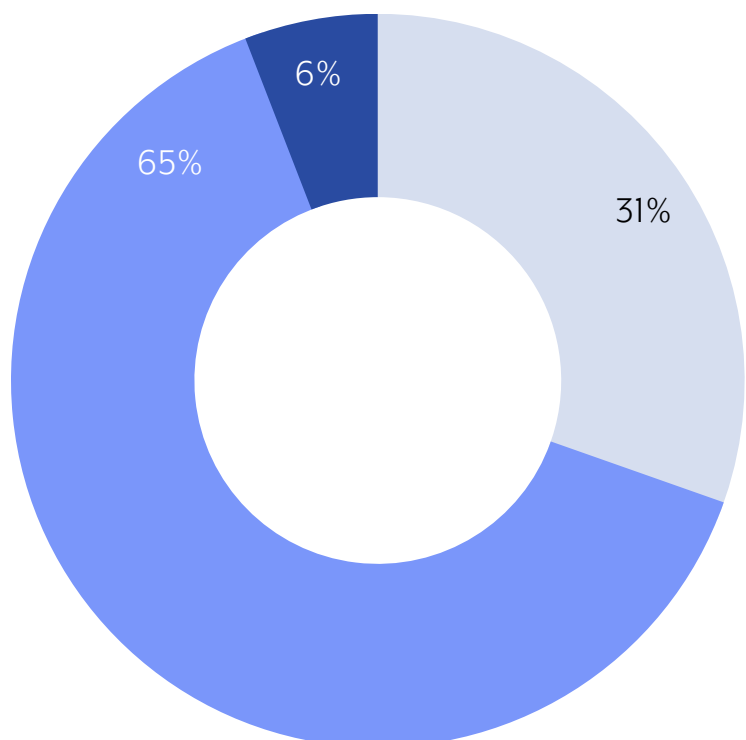
YOUTUBE

N = 237



TIKTOK

N = 194



*Note: Percentages may not add up to 100% as some videos contain multiple forms of harmful content.

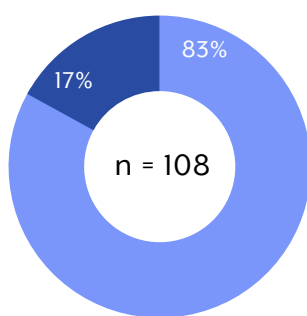
FINDINGS

Demographics

YOUTUBE

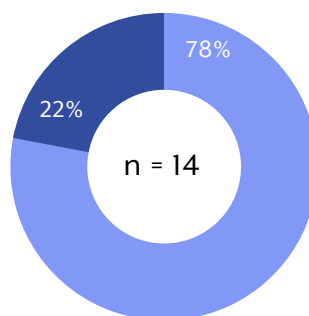
Unique Users (N) = 205

GENDER
(Self-Reported + Estimated)



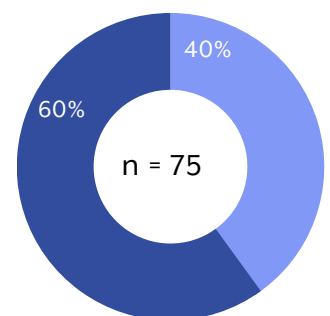
Female
Male

AGE
(Self-Reported)



Below 18¹
Above 18²

AGE
(Estimated)

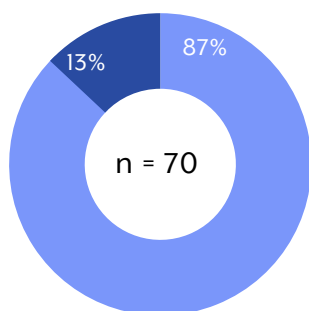


Young Adult
Below 18

TIKTOK

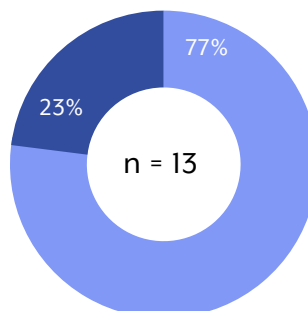
Unique Users (N) = 136

GENDER
(Self-Reported + Estimated)



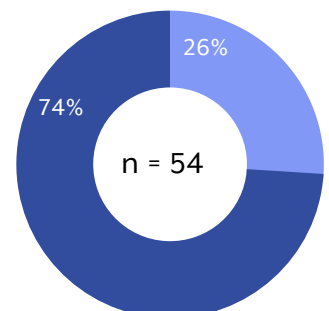
Female
Male

AGE
(Self-Reported)



Below 18³
Above 18⁴

AGE
(Estimated)



Young Adult
Below 18

¹ YouTube users' self-reported age (Below 18): 12 Y/O (n = 1), 13 Y/O (n = 3), 15 Y/O (n = 5), 16 Y/O (n = 2)

² YouTube users' self-reported age (Above 18): 19 Y/O (n = 1), 20 Y/O (n = 1), 24 Y/O (n = 1)

³ TikTok users' self-reported age (Below 18): 14 Y/O (n = 1), 15 Y/O (n = 4), 16 Y/O (n = 2), 17 Y/O (n = 3)

⁴ TikTok users' self-reported age (Above 18): 19 Y/O (n = 1), 20 Y/O (n = 1), 21 Y/O (n = 1)

HASHTAG GENERATION TECHNIQUES

Utilising hashtags enables discovery and propagation of content across social media platforms. For harmful content, the hashtags adopted often serve as coded language shared among a community of like-minded users to circumvent content moderation on social media platforms. Crafting hashtags can be a highly sophisticated process – or perhaps, considered an art – and requires an appreciation of nuances. Broadly, there are four techniques to hashtag generation (see Table 2):

Table 2

Hashtag Generation Techniques

Hashtags

#tw | #sh | #ana

#suwerslide |
#depreshun

#thinspo

#ednotsheeran

Technique

Abbreviations and acronyms

E.g., "tw" for trigger warning, "sh" for self-harm, "ana" for anorexia

Deliberate misspelling of terms; substitution with phonetically similar words

E.g., "suwerslide" to imply "suicide" and "depreshun" to imply "depression"

Euphemisms

E.g., "thinspo" as a shorthand for "thinspiration"

Hijacking famous subjects

E.g., hijacking popular singer Ed Sheeran's name to signify eating disorders as the shorthand for eating disorders is "ed"

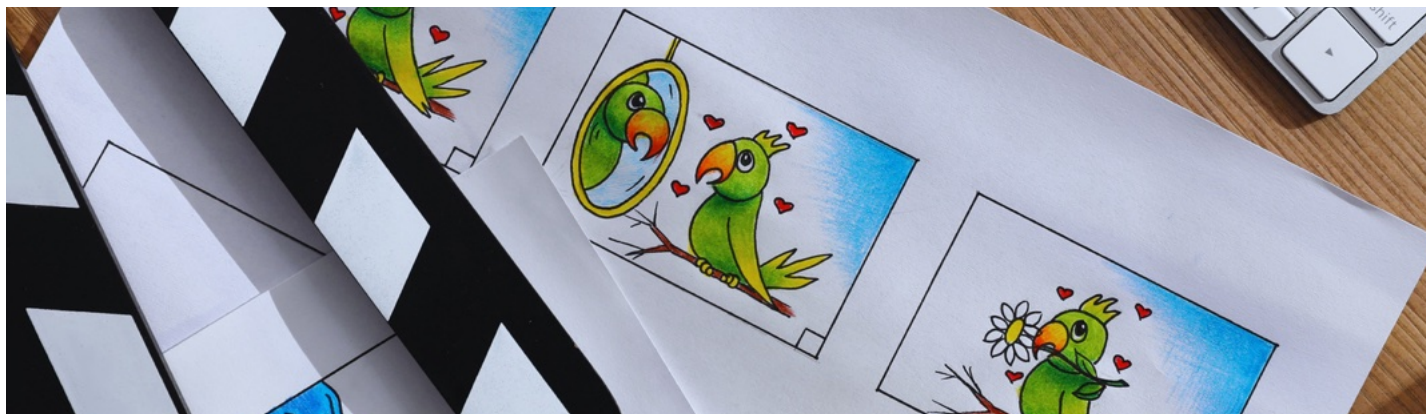
CONTENT CURATION TECHNIQUES

We now present a list of common content curation techniques that users employ to conceal their harmful content. It is crucial to note that these techniques should not be viewed in isolation as users tend to skilfully employ a combination of techniques listed below to fashion their content.

To protect the users' privacy, we deliberately left out screenshots and links to the social media posts and rely on writing to describe what we found. We would like to provide a content disclaimer to warn our readers that there is a possibility of finding offensive content in the examples included in the work presented.

YOUTUBE

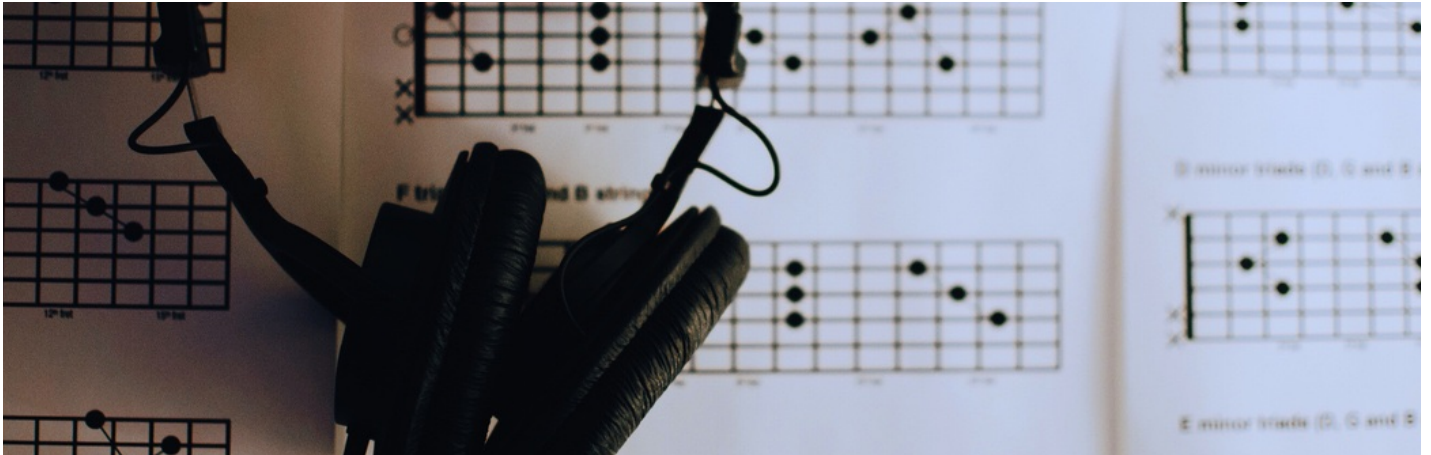
Animations | 31%



Animations are more prominent in content pertaining to self-harm and suicide. The animations are not professional in nature, but the creation of such content certainly requires more skill and expertise.

The animations generally make explicit references to self-harm or suicide, such as using a knife to slash one's arm leaving behind red markings and bands to signify bleeding, jumping off the rooftop of a building or even an animated character placing a gun to its head and firing off. They may be paired with music containing sad lyrics to amplify the pain and sadness.

Music | 30%



Music is primarily used in content relating to self-harm and suicide. These take two forms: (a) selective inclusion/omission of verses of a song which contains sad lyrics; or (b) distortion of selected verses of a song by accelerating or slowing it down.

>> For content relating to [self-harm](#), pertinent examples include:

"Fat Funny Friend" by Maddie Zahm

*"But my effort's in vain.
They can't relate to how I've drawn out the Sharpie where I take the scissors.
If that's what it took for me to look in the mirror
I've done every diet to make me look thinner
So why do I still feel so goddamn inferior?"*

Although this song recounts Zahm's personal experience as being the fat friend of the group, it has been used to convey self-harm behaviours. The following verses are used to signify self-cutting and the emotional anguish that the user is experiencing:

"Alien Blues" by Vundabar

*"My teeth are yellow, hello world
Would you like me a little better if they were white like yours?
I need to purge my urges, shame, shame, shame"*

The following verses were sped up to create completely different sounding lyrics and used in the context of self-mutilation – more specifically, when users are “caught” by their friends or family for having engaged in self-mutilation:

>> For content relating to [suicide](#), pertinent examples include:

"My R" by Annapantsu (English Cover by Lollia)

*"There's no one here today, I guess it's time.
It's just me, myself and I
There is no one who can interfere, no one to get in my way here
Taking off my yellow cardigan, watching my braids all come undone.
This petite girl, short as can be, is gonna jump now and be free"*

The song showcases one's internal struggle with their suicidal intent, and the following verses were adapted by users to express a desire to jump off a building:

"自傷無色Self-Inflicted Achromatic" by Hatsune Miku (English Cover by JubyPhonic)

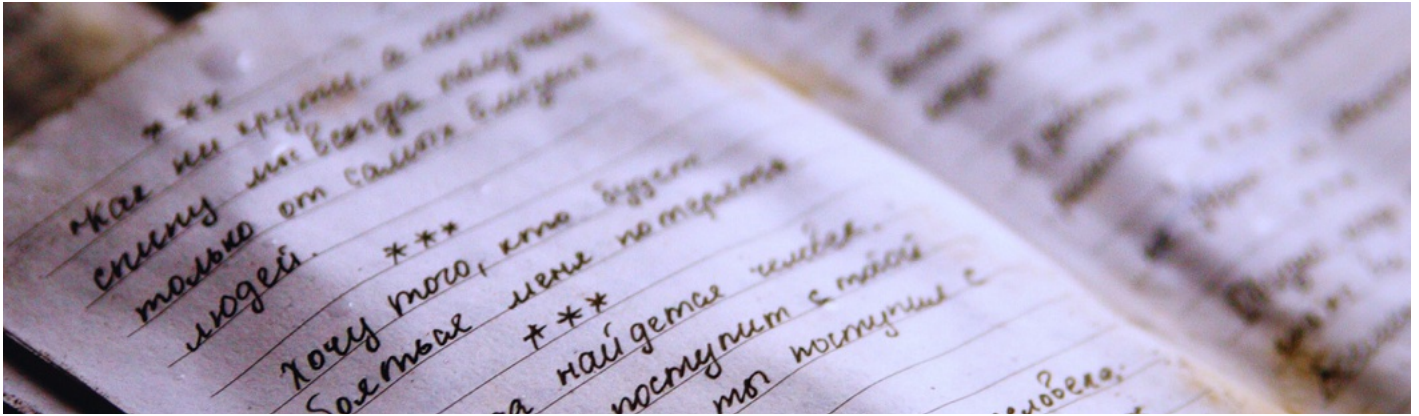
*"Now as I am, I understand it's best I die and soon.
Just by living, I'm hurting them another day
Hundreds cry, all I do is ruin everything
Nobody wanted me, no one there to need
If only I could live in that kind of world I dreamed"*

The English cover of the song reflects the journey of someone who had masked their depression, questioned their worth and experienced suicidal tendencies. The following verses were adapted by users to express their desire to "die" as they are hurting badly:

Separately, there were also a few videos which feature full albums of suicidal black metal music. In these instances, the album covers on these videos feature illustrations of knives and scissors to imply self-cutting, and ropes to signify suicide by hanging. The vocals are primarily strained screams with undiscernible lyrics, but unmistakably expresses suffering. Topped with the electric guitar sounds and deep bass, the listening experience may be described as heavy, chaotic and torturing.

There were also standalone music pieces with lyrics conveying self-harm and suicidal intent. For such pieces, there were no formal album covers or illustrations. The vocals are muffled, low-pitched with a relatively monotonous tone which contributes to a downcast listening experience.

Diaries and Vlogs | 23%



These are particularly prominent for content pertaining to eating disorders. Users often offer a trigger warning before the share what they eat within a day or across a specific time period. Users also engage in calorie counting where they note down the calories consumed, calories burned, and their net caloric gain/loss.

Some also choose to include a record about their current weight, interim weight goal and their ultimate weight goal. Low-diet promotion may be difficult to detect, especially if the users do not show actual numbers of their caloric intake and loss.

Video Cuts/ Voiceovers of TV Shows | 11%



This technique entails using video cuts and/or audio voiceovers taken from TV shows. On YouTube, this was used to present self-harm behaviours, particularly burning and cutting. Users would then include additional texts and images to better contextualise their experiences.

>> *Examples of shows and movies adapted by users include:*

Ginny & Georgia

There were two scenes which were adapted by users: (a) Ginny burns herself with a lighter; and (b) Georgia find out that Ginny burns herself and confronts the latter. Users tapped on these scenes and offered accompanying texts relating to their personal experiences of self-harm, or images and videos showcasing their scars.

Shameless

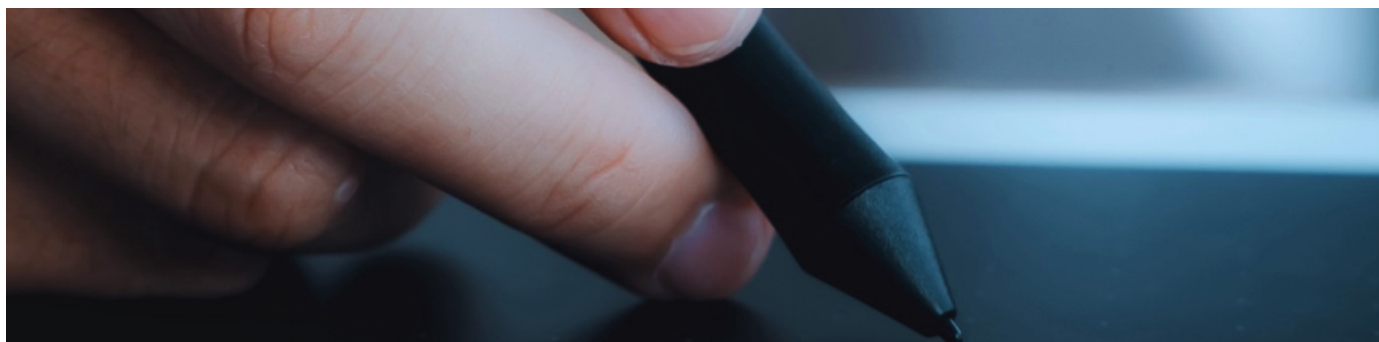
One scene where Fiona was engaged in an argument with Frank, and she yelled: “I was nine! Nine, and taking care of you” was adapted by users. In adapting this scene, users included texts to “modify” the scene and imply that they had engaged in self-harm behaviours since they were minors.

Sketches | 9%



Sketches tend to be observed in content pertaining to self-harm and suicide. These take the form simple sketches of the human figure, coupled with knives and/or ropes. These sketches may also be animated. The colour scheme of such content tends to be monochromatic, except for the use of bright red markings or streaks to signify blood and self-mutilation.

Coded Language | 8%



Coded language - purely in textual format - was featured in self-harm content, particularly in videos about self-mutilation. The utilised texts are non-offensive when viewed standalone, but in the context of self-harm, the unfolding dialogue takes on an implicit meaning that the users have been (or enjoy) cutting themselves.

The following is an exemplar dialogue visible in multiple videos:

“Person 1: I like cutting flowers.

Person 2: I like cutting paper.

Person 3: I like cutting hair.

Me: ...”

There were a few videos which conveyed more explicit intentions through these expressions:

“I cut my hair, I cut paper, I want to cut more than that”

“I need to stop paper cutting everyday”

“I like to scratch mys-... dog”

Non-Human Characters | 2%



The use of non-human characters was observed in a small number of self-harm content, specifically for self-mutilation. Sketches or lo-fi animations of creatures including a goat, cat and wolf slashing themselves using blades or their own nails and bleeding profusely were observed.

Game References | 2%



Game references was featured in a small number of self-harm content, specifically for self-mutilation. The game “Fruit Ninja” developed by Halfbrick Studios – where users slice fruit thrown into the air by swiping and slashing fruit to reach a high score – were referenced in all observed instances. This was accompanied by a text saying encouraging others to “download Fruit Ninja” (i.e., implicitly suggesting to slash oneself) as it is “fun” or “amazing”.

TIKTOK

Coded Language and Images | 44%



Featured across all types of harmful content. The user often uses language (in the captions, within the video, or both) and images which are non-offensive when viewed standalone. However, when viewed concurrently with a contextual lens, an underlying message with darker undertones begins to surface (i.e., implicit messages).

Here are two pertinent examples to illustrate this point:

- A video showcases two images – the first depicting an arm with a broken pink heart emoticon, the second depicting a zebra with a bandaged red heart emoticon. Viewed individually, nothing appears out of the ordinary. However, when interpreted in context with the hashtags “#TWSH” and “#vent” embedded in the post, the underlying message is about self-harm behaviours. Specifically, the relief which the user experience from cutting themselves (i.e., the zebra signifies the scars from cutting, and the bandaged heart signifies emotional relief).
- A video showcases an image of laxatives with the overlaying words saying: “Yes, you are my lover”. The first transition occurs in which a heart-shaped animation envelopes the screen and a new set of words emerges: “There is no other”. A final transition occurs in which the heart-shaped animation envelopes the screen once again before showing an animated human figure smiling while sitting on a toilet bowl. Without further context, it could be interpreted as relief from constipation. However, when interpreted in context with the hashtags “#ednotsheeran” embedded within the post, the underlying message is about eating disorders. Specifically, encouraging laxative abuse which can cause life-threatening conditions.

Music | 24%



The use of music is a common technique users adopt to convey their underlying feelings and intentions, and fit the narrative users are portraying regarding eating disorders, self-harm and suicide.

The use of music on TikTok tends to take two forms – (a) selective inclusion/omission of verses of a song which contains lyrics which aligns with the users’ intentions; and (b) selective inclusion/omission of verses of a song paired with visual texts to “modify” the lyrics of the original song.

>> For content relating to [eating disorders](#) a pertinent example is:

"Prom Queen" by Beach Bunny

*"Shut up, count your calories
I never looked good in mom jeans
...
Maybe I should try harder
You should lower your expectations
I'm no quick-curl barbie
I was never cut out for prom queen"*

The song expresses one's bodily insecurities and the desire to achieve the perfect body and ideal beauty standards. The following verses were adapted to reflect the users' disappointment in their bingeing habits:

>> For content relating to [self-harm](#) a pertinent example is:

"Dollhouse" by Melanie Martinez

*"No one never listens
This wallpaper glistens
Don't let them see what goes down in the kitchen
...
Picture, picture, smile for the picture
Pose with your brother, won't you be a good sister?
Everyone thinks that we're perfect
Please don't let them look through the curtains"*

Users adapted the following verses of the original music and then included accompanying texts on the videos to "modify" the lyrics and confess their inner thoughts and personal circumstances:

Note:

In the accompanying text observed on one video, the user conveys a message suggesting that not everything observed on the "picture" is as "perfect" as it seems, and that they wish to hide their scars under the "curtains"

>> For content relating to [suicide](#), a pertinent example is:

"Consume" by Chase Atlantic (ft. Goon Des Garçons)

*"These voices in my head screaming, "Run now"
I'm praying that they're human
...
Please understand that I'm trying my hardest
My head's a mess but I'm trying regardless"*

Similar to self-harm content, users included texts within the video to “modify” the lyrics to express their inner thoughts:

Note:

In the accompanying text observed on one video, the user notes that they are hearing “voices” telling them to kill themselves, and that they are “praying” it would go away

Diaries and Vlogs | 20%



Most prominent in content about eating disorders. Similar to the techniques described in the section on YouTube above, users often engage in calorie counting – noting the calories consumed and burned, and net caloric gain/loss across a period of time. Users also tend to include a record about their current weight, interim weight goal and their ultimate weight goal.

Video Cuts/Voiceovers of TV Shows/Movies | 13%



Similar to the earlier discussion on YouTube, users may include video cuts and/or audio voiceovers taken from TV shows and movies. This involves including video snippets or solely extracting the audio recordings from the TV shows and/or movies before pairing it with texts to draw references to their personal situation. On TikTok, however, such techniques were primarily used to portray eating disorders.

An example of a TV show is:

"Insatiable"

Multiple cutscenes showing Patty gorging on junk food and crying were adapted by users to portray their suffering from binge-eating disorder and subsequently, venting about their inability to lose weight.

Celebrity Images/Videos | 12%

This technique was most pertinent in content about eating disorders. Users would include images and/or videos of celebrities in their content about anorexia, binge eating and purging.

Interestingly, the celebrities featured are primarily female Korean idols who are generally perceived to be thin and representative of the K-Pop beauty “standards”. A few users also implied their intentions to reach these celebrities’ profile weights (which has been denounced as unrealistic by observers).



Jang Wonyoung [I*ZONE/IVE]



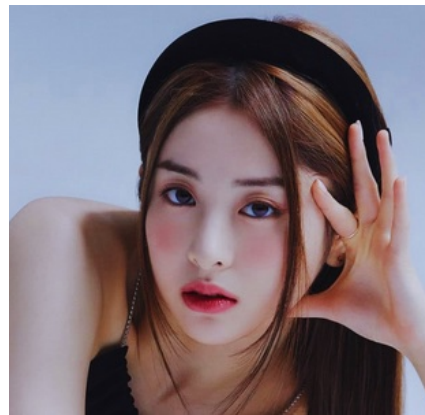
Kang Hyewon [I*ZONE]



Sana [TWICE]



**Jisoo, Jennie, Lisa, Rosé
[BLACKPINK]**



Yunjin [LE SSERAFIM]



Yuna [ITZY]



Minji, Hyerin [NewJeans]



Bahiyyih, Yujin [Kep1er]



Miyeon [(G)I-DLE]

Photos taken from: Wikipedia, Kpopping, Newstap, Pinterest, Korea.net and Kpopn

Special Characters and Symbols | 11%



By employing lookalike characters – including variations of the Latin alphabet and/or special symbols – that bear resemblance to the letters of specific words, users can convey their intended meaning while evading detection by the platform’s automatic filters and moderators when posting harmful content.

A list of words that were observed is appended below for reference:

Eating Disorders

eátiñg dĩ\$order\$
4na
proama
an0r3x!c
@norex!a
bü!m!c
st@rve

Self-Harm

cv+s
dangeour
zcars
\$h

Suicide

kllling
KY\$
d!€

LIMITATIONS AND IMPLICATIONS

From this research exercise, we noticed that users who post harmful content often utilise a range of keywords and hashtags to help categorise their content. Arguably, this could be to increase social media reach as keywords and hashtags allow content discovery through searches on the platform.

While this allows us to efficiently conduct data extraction for analysis, the eventual data collected will still be limited in scope given the innovativeness of social media users to permutate and distort existing words or phrases (or even coin new ones), which are coloured with darker undertones. Furthermore, given the limited time, we could only scan and extract a limited volume of data points.

Our research findings do not fully represent the scale of harmful content circulating on social media platforms, but it certainly suggests that searching for and accessing harmful content is [not challenging at all](#).

This is of concern as the vicarious consumption of such harmful content may create a false impression of a desirable identity, glamorise such behaviours, or worse, inspire and encourage the replication of such behaviours.

Undoubtedly, the ramifications for both physical and mental well-being of those being influenced are profound.

Separately, we cannot affirm that users who post harmful content will always utilise keywords and/or hashtags. This presents a problem for both researchers who wish to examine the extent of circulation of harmful content on platforms, and for the platforms who engage in content moderation. This problem is further compounded as users are adopting creative techniques adopted by users to mask harmful content and evade platform policing.

Nonetheless, our writing offers a promising start into examining the (ever-evolving) variety of content curation techniques to disguise harmful content. Future monitoring needs to use multi-modal detection such as image, sound, and video detection technologies to search through the social media content.

This signals a need to devise novel content detection techniques which take advantage of technology to counter the volume and velocity of content, but heavily guided by human interventions to ensure contextual appreciation across territories and cultures to stay ahead of the curve.

RECOMMENDATIONS

To the Government

Third-party content monitoring is needed. The monitoring body needs to be independent to maintain its integrity. A multi-stakeholder approach is necessary in assembling and operating this monitoring body. The monitoring results serve as an alternative source of evidence, not as a replacement of platforms' self-reports.

To the Platforms

Content moderation requires human engagement. Human moderators are trained to make timely decisions and be protected from work injuries. Communities, especially those of vulnerable user groups, have to be included in the entire process of preventing and countering online harms, from the early stage of algorithm and features design, to the late stage of case appeals. The consequences of platforms not being able to fulfil these responsibilities need to be made clear.

To Parents and Educators

Education of parents and educators regarding the issue needs to be made easily accessible. Social services that support parents and educators need to be made widely available. Both the monitoring body and platforms' internal process should be open to these actors.

To Youth and other Vulnerable Users

Self-expression can be done by making the content private or non-searchable. Social media platforms are not always the best space for getting oneself understood. Protect yourself. Protect others.

Acknowledgments

We would like to thank the following individuals for the engaging and thought-provoking discussion on the report's preliminary findings during the roundtable event:

Anita Low
(TOUCH)

Huan Ting Lee
(MCI)

Arun Elangovan
(Isentia)

Jenna Wang
(Isentia)

Benjamin Goh
(Independent)

Keunying Hur
(Isentia)

Grace Ng
(MCI)

Natalie Pang
(NUS)

We would also like to express our appreciation to the following individuals for their feedback on the report:

Anbar Jayadi

Candera Chan

Enrico Tajanlangit

Nathaniel Ong

(TikTok)