Hidden in Plain Sight: Exploring the Intersections of Mental Health, Eating Disorders, and Content Moderation on TikTok

Charles Bickham[*1,2], **Kia Kazemi-Nia**[*1,2, 3], **Luca Luceri**[2], **Kristina Lerman**[1, 2], **Emilio Ferrara**[1,2]

[1] Department of Computer Science, Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA
[2]Information Sciences Institute, University of Southern California, Marina Del Rey, CA, USA
[3]Department of Psychology, Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA, USA
cbickham@usc.edu, kazemini@usc.edu, lluceri@isi.edu, lerman@isi.edu, emiliofe@usc.edu

Abstract

Social media platforms actively moderate content glorifying harmful behaviors like eating disorders, which include anorexia and bulimia. However, users have adapted to evade moderation by using coded hashtags. Our study investigates the prevalence of moderation evaders on the popular social media platform TikTok and contrasts their use and emotional valence with mainstream hashtags. We notice that moderation evaders and mainstream hashtags appear together, indicating that vulnerable users might inadvertently encounter harmful content even when searching for mainstream terms. Additionally, through an analysis of emotional expressions in video descriptions and comments, we find that mainstream hashtags generally promote positive engagement, while moderation evaders evoke a wider range of emotions, including heightened negativity. These findings provide valuable insights for content creators, platform moderation efforts, and interventions aimed at cultivating a supportive online environment for discussions on mental health and eating disorders.

Warning: This paper discusses eating disorders, which some may find distressing.

Introduction

This study explores the interplay between mental health and social media, using a case study of TikTok and eating disorders. Eating disorders represent a complex mental health condition characterized by disruptions in eating and related behaviors. These conditions, which include anorexia, bulimia, and binge eating disorder (BED) (Walsh, Attia, and Glasofer 2020), have alarming mental health impacts: one study reported that up to 23% of individuals diagnosed with BED in the US had attempted suicide, with 94% experiencing a lifetime of mental health problems (Keski-Rahkonen 2021). The COVID-19 pandemic further complicated the landscape of eating disorders (Hogue and Mills 2019; Rodgers et al. 2020; Schlegl et al. 2020). During this period, there was a notable surge in symptomatology, accompanied by diminished access to treatment for individuals struggling with eating disorders (Cooper et al. 2022). This crisis highlighted the urgency of understanding the dynamic interplay between external stressors, mental health, and the manifestation of eating disorders.

TikTok, a platform for sharing short-form video content, has billions of users worldwide (Iqbal 2021) and is especially popular among adolescents. While TikTok provides a creative outlet for many young people, it has also become a space for discussions about mental health and eating disorders (Herrick, Hallward, and Duncan 2021; McCashin and Murphy 2023). The brevity and immediacy of TikTok videos, often accompanied by succinct descriptions containing hashtags, present an opportunity to investigate how individuals engage with emotional issues, as well as broader questions about the role of social media in mental wellbeing, particularly in adolescents (Frieiro Padin et al. 2021; Pruccoli et al. 2022; Sha and Dong 2021).

This work explores the emotional landscape of TikTok videos related to eating disorders, focusing on the role of hashtags in the discoverability and organization of content. While TikTok heavily moderates searches for harmful content that promotes or glorifies eating disorders (Casilli, Pailler, and Tubaro 2013), users have adapted to evade moderation through the use of coded hashtags (Herrick, Hallward, and Duncan 2021; Cobb 2017). A study of TikTok by the Center for Countering Digital Hate (CCDH)[1] found content promoting eating disorders relied on coded hashtags to avoid getting flagged by content moderation algorithms. These coded hashtags were based on misspellings, abbreviations, or references to the artist Ed Sheeran, whose name coincidentally begins with "ED" – an abbreviation commonly associated with eating disorders. For the purpose of this study, we refer to such hashtags as *moderation evaders*. In this paper, we analyze the emotional expressions in content across different hashtags pertaining to eating disorders, including moderation evaders and mainstream hashtags.

- **RQ1:** Do moderation evaders co-occur with mainstream hashtags related to mental health and healthy living?
- **RQ2:** How do emotional expressions in TikTok video descriptions differ between content associated with mainstream hashtags and moderation evaders?
- **RQ3:** How do emotional expressions in TikTok video descriptions correlate with those in user comments, and

[*] These authors contributed equally.

[1] https://counterhate.com/wp-content/uploads/2022/12/CCDH-Deadly-by-Design_120922.pdf#page=44.12

how does this interaction differ across mainstream hashtags and moderation evaders?

Our study of the emotional expressions within video descriptions and user comments reveals systematic differences between mainstream hashtags and moderation evaders. While mainstream hashtags tend to promote an overall positive engagement, moderation evaders evoke a broader range of emotions, including heightened negativity. The disparity in emotional engagement suggests that moderation evading hashtags are used for spreading problematic content related to eating disorders. Moreover, since moderation evaders cooccur frequently with mainstream hashtags, this raises concerns about the potential exposure of users to harmful content, as moderation evaders often circumvent platform regulations.

Our findings underscore the complex emotional landscape of TikTok content related to eating disorders and emphasize the need for tailored moderation strategies and interventions to cultivate a supportive online environment conducive to discussions on mental health and eating disorders. In the subsequent sections, we present related works, our methodology, discuss our results, and draw meaningful conclusions that contribute to the evolving discourse on mental health and eating disorders in the digital age.

Related Works

Social Media and Body Image Concerns

The rise in social media use has fueled worries about its impact on negative body image concerns. Social comparison has been found to play a negative role in how people think about their body image (Hülsing 2021; Mink and Szymanski 2022; Westenberg and Oberle 2023; Jiotsa et al. 2021; Festinger 1957). Women and young girls especially tend to suffer from negative body image concerns (Peng 2023; Liu 2022; Hogue and Mills 2019). Social media platforms that are based on photos are thought of to be more negative when it comes to body image since they tend to be more focused on physical appearance (Karsay et al. 2021; Rodgers and Rousseau 2022). With photo-based platforms, there are more opportunities for people, especially women, to self-objectify, internalize appearance ideals, and compare themselves negatively. This can lead to many mental health risks that include but is not limited to unipolar depression, sexual dysfunction, and eating disorders (Fredrickson and Roberts 1997). Additionally, the use of photo-based platforms allows for images to be manipulated. Studies show that there has been a link between lower body image and self-esteem based on posting exposure to manipulated images (Chua and Chang 2016; Wick and Keel 2020; Cohen, Newton-John, and Slater 2018; Kleemans et al. 2018). Also, many people may not be aware of the images being manipulated, which can lead to a normalization of unrealistic body and beauty ideals (Marks, De Foe, and Collett 2020). Exposure to the thin ideal has been linked to an increase in body dissatisfaction, eating disorder symptoms, and negative mood in women (Hawkins et al. 2004; Fardouly et al. 2015; Tiggemann, Anderberg, and Brown 2020). The presence of a negative self-body image in an individual is identified as a contributing factor to eating disorders (Manaf, Saravanan, and Zuhrah 2016).

Social Media and Eating Disorders

Social media use has been linked with distorted eating over recent years Holland and Tiggemann (2016); Zhang et al. (2021). With the rise of this trend, there have been efforts to prevent eating disorders. (de Valle and Wade 2022) showed that self-criticism intervention can be a strategy to address this need. Communication strategies have also been looked into to help the prevention of eating disorders (Rando-Cueto, de las Heras-Pedrosa, and Paniagua-Rojano 2023). Lerman et al. (2023) map the relationship between social media and eating disorders to an online radicalization process (Kruglanski et al. 2014; Wang et al. 2023). Exploring the concept of online radicalization within “pro-ana” communities, (Lerman et al. 2023) highlights how social media platforms facilitate radicalized behaviors by creating echo chambers that can normalize distorted eating. The study emphasizes the importance of understanding and quantifying the impact of these online communities to develop strategies aimed at promoting better mental health. Branley and Covey (2017) details how easy it is for someone to “stumble” upon potentially harmful ED content without explicitly searching for it.

Specifically on TikTok, Herrick, Hallward, and Duncan (2021) analyzes the mainstream hashtag #edrecovery, while (Greene and Norling 2023) analyzes the textual content, using mixed methods, on the hashtag #bedrecovery. (Hung et al. 2022) conducts content analysis on the hashtags #fitspiration and #thinspiration. Greene et al. (2023) performed a comparative analysis on pro-recovery communities across five eating disorder hashtags: #anarecovery, #arfidrecovery, #bedrecovery, #miarecovery, and #orthorexiarecovery. (Dondzilo, Rodgers, and Dietel 2023) supports indirect connections between involvement with appearance/eating-related content on social media and eating disorder symptoms, with mediation through elevated exposure to recommended content in this category and increased levels of upward social media appearance comparisons.

The impact of social media and eating disorders on the youth have also been studied (Lonergan et al. 2020; Pruccoli et al. 2022; Corzine and Harrison 2023; Salomon 2020). These studies emphasize the need for an understanding of the relationship between social media use and eating disorders among young adults and children.

Methodology & Data

Dataset

For the purpose of this research, we collected information about TikTok videos focusing on content associated with various issues related to body image, dieting, and eating disorders. To collect the data, we curated a set of hashtags that reflect the diversity of topics within the TikTok platform. See the Appendix for the full list of keywords. Spanning the timeframe from December 2016 to April 2023, the dataset has a total of 14,816 posts and 562,856 comments associated with these videos.

Cluster	Count	Average Likes	Average Shares	Average Plays	Average Comments
ED recovery	182	239,976.80	4,915.86	1,948,781.66	1,868.81
Healthy living	80	149,775.51	3,299.18	1,399,869.32	1,317.07
Body positivity/acceptance	31	31,887.22	853.84	353,926.05	276.77
Juice-related content	35	22,947.38	752.35	233,925.57	243.05
Beauty and body positivity	81	296,309.77	6,105.77	2,491,468.19	2,327.29
Miscellaneous	86	317,098.94	4,465.23	2,417,362.08	2,637.50
Music promotion	38	25,090.83	875.28	282,339.19	227.03

Table 1: Statistics for each hashtag community

The dataset encompasses various aspects of TikTok videos with each entry containing textual information. Textual information is captured through the video description (Description) and a list of hashtags used in the video (Challenges). Additionally, the dataset includes the comments the videos received.

Hashtag Co-Occurrence Graph

A hashtag co-occurrence graph operates as a network that outlines the relationships among hashtags present in a series of social media posts. When two hashtags share an edge, it signifies their joint appearance in a post, and the weight of this edge reflects the frequency of these occurrences—essentially measuring how often they appeared together in our dataset.

Measuring Emotions

Textual content conveys signals related to emotions and feelings, encompassing both positive sentiments, such as joy and love, and negative ones like anger and disgust. In our approach, we employ an emotion detection tool inspired by SpanEmo (Alhuzali and Ananiadou 2021), named Demultiplexer (Demux) (Chochlakis et al. 2023). The tool takes the categories (in this case, emotions) as the first input sequence and the actual content as the second sequence. The contextual embeddings specific to each emotion contribute to deriving probabilities for individual emotions. Our application of Demux extends to every video description and comment. The emotions included were anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust, and none. For the purpose of this study, anticipation, pessimism, surprise, and trust were grouped into a single category named "other emotions" similar to (Lerman et al. 2023).

Results

The methods detailed above were used to analyze our dataset and answer our research questions about the co-occurrences and emotional patterns across mainstream hashtags and moderation evaders.

RQ1: Co-Occurrence of Moderation Evaders and Mainstream Hashtags

To address our first research question, we created a hashtag co-occurrence graph of popular hashtags (that occurred at least 30 times), with 612 hashtags in total. Figure 1 shows the hashtag co-occurrence network. The node size in this graph corresponds to the hashtag's PageRank centrality. We used the Louvain algorithm to identify clusters of highly interlinked nodes, revealing seven main communities. The central hashtags within these clusters are listed below, including those that are intentionally misspelled.

- Eating disorders (ED) recovery community with hashtags #edrecovery, #edrecvery (*sic*), #anorexiarecovery, . . .
- Healthy living community: #exercise, #workout, #healthyrecipes, . . .
- Body positivity/acceptance community: #bodypositivity, #fatacceptance, #selflove, . . .
- Juice-related content: #juicerecipes, #healthyjuice, #juicingforhealth, . . .
- Beauty and body positivity community: #thighs, #beautiful, #curvy, . . .
- Music promotion: JuiceWRLD-related content community which include #juicewrld #juicewrldmusic, . . .
- Miscellaneous: #water #aquarium

The largest cluster in this network, with 182 hashtags and over 1.9 million views per video, is devoted to *ED recovery*. Table 1 gives a full breakdown of the statistics for each cluster.

We identified 10 moderation evaders within the 612 hashtags, and all of them were located in the ED recovery community. These hashtags were: #edrecocery, #edsheeranrecoveryy, #edawarewness, #anarecvery, #edrecov, #anoreksja, #edtt, #edsheeran, #ednotsheeren, and #ana. Moderation evaders frequently co-occurred with mainstream hashtags like #fyp, #fearfood, #recovery, #ed, and #mentalhealthmatters. For instance, #edrecocery appeared 138 times in total, with 64 occurrences alongside #ed, while #anarecovry appeared 61 times in total, with 55 occurrences with #recovery. See table 2 for full list. The co-occurrence of mainstream hashtags with moderation evaders suggests that users who search for terms such as #ed or #recovery could be exposed to content containing moderation evaders. The hashtag #anoreksja also co-occurred with #tw (trigger warning) over 20 times. Furthermore, the hashtag #wl (weight loss) co-occurred with #edtt and #ednotsheeren over 10 times. The associations of moderation evaders with *weight loss* (#wl) and *trigger warnings* (#tw) indicate potential triggering content that may be harmful to individuals struggling with eating disorders.

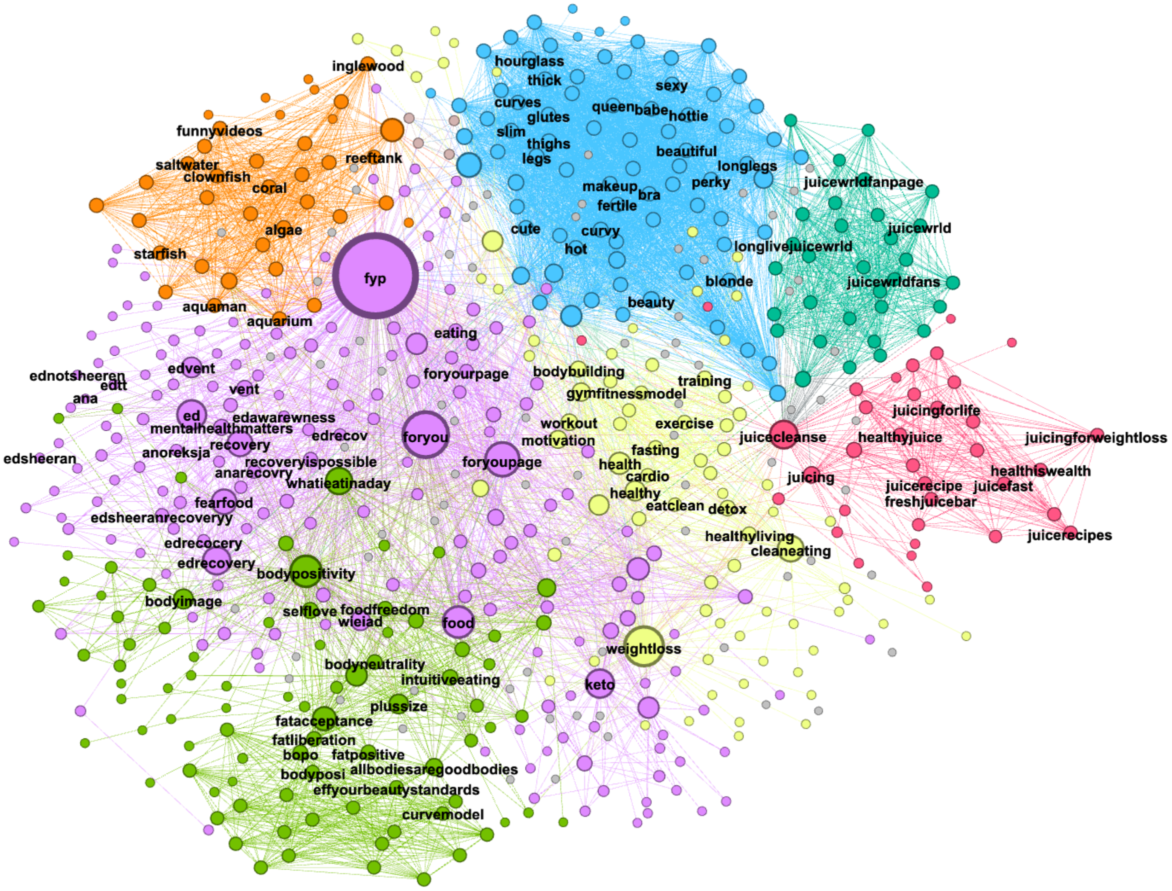


Figure 1: Purple = ED Recovery, Yellow = Healthy Living, Green = Body Positivity/Acceptance, Pink = Juice-Related, Blue = Beauty and Body Positivity, Teal = Music Promotion, Orange = Miscellaneous.

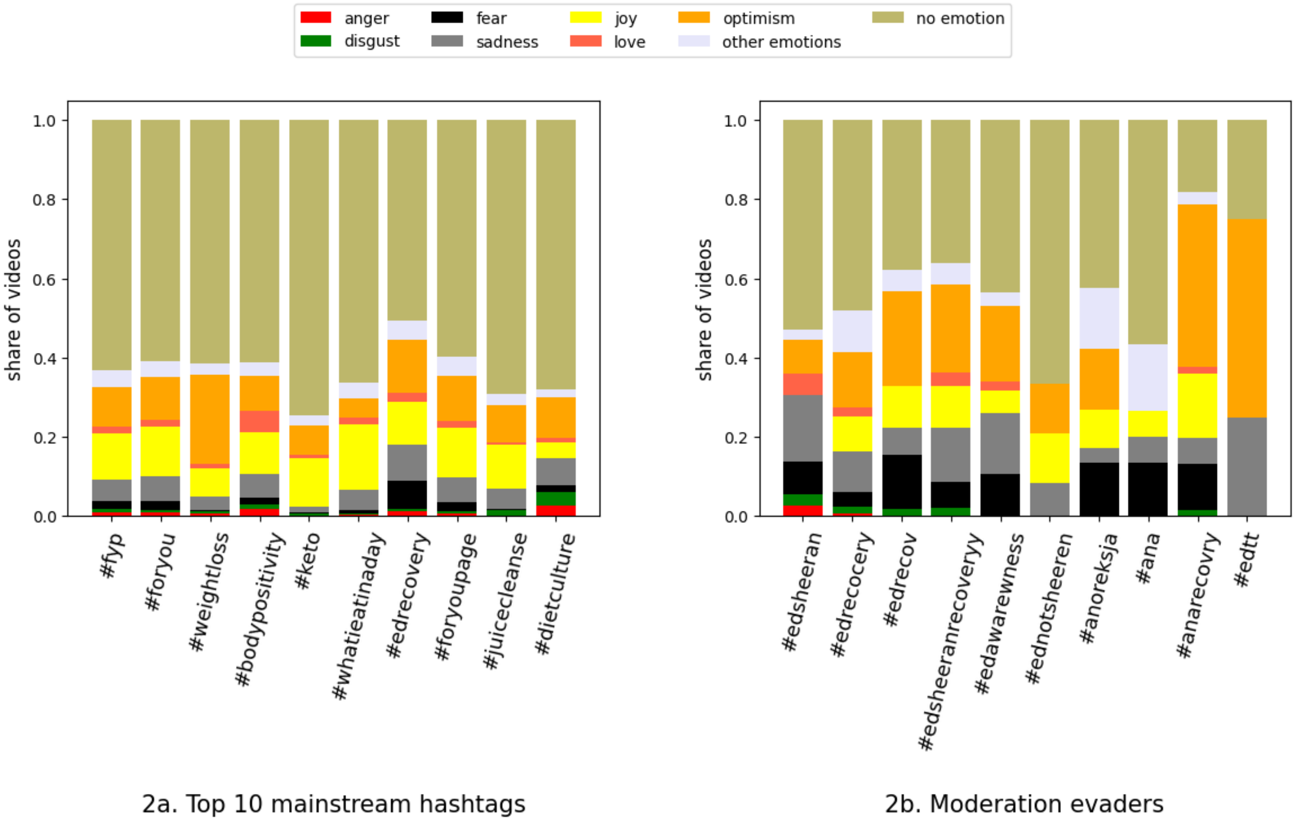
Findings and Remarks: Addressing RQ1, the analysis of TikTok hashtags revealed a prominent cluster related to ED recovery, with 182 hashtags with over 1.9 million views per video. This contrasts with Twitter, where harmful content promoting eating disorders is far more common (Lerman et al. 2023). While this shows that TikTok provides a more positive, pro-recovery platform, the presence of moderation evaders within this cluster suggests that harmful content may be intentionally obscured to avoid moderation. Moderation evaders are associated with hashtags weight loss (#wl) and trigger warnings (#tw), highlighting links to potentially problematic content in the context of eating disorders. Moreover, the fact that these moderation evaders are linked to mainstream hashtags, such as #fyp and #recovery, suggests potential exposure to pro-eating disorder content for users searching for recovery-related content.

RQ2: Emotions in Video Descriptions

To address our second research question, we compared emotional expressions in the descriptions of TikTok videos tagged with mainstream hashtags and moderation evaders. We first focused on videos that have been tagged with any of the ten most popular mainstream hashtags (based on the number of occurrences). Joy consistently emerged as the dominant emotion, followed closely by optimism. Among negative emotions, sadness was the most common. A substantial portion of posts expressed "no emotion", potentially influenced by the succinct and hashtag-centric nature of TikTok video descriptions. See Figure 2a for the emotional analysis results for the mainstream hashtags.

Emotional analysis of the descriptions of videos tagged with any of the ten moderation evaders that we previously identified revealed them to be more emotional overall. Optimism was the dominant emotion. Sadness and fear were also higher, suggesting a more negative emotional tone (Figure 2b).

Comparing emotional expressions of videos tagged with mainstream hashtags and moderation evaders revealed interesting patterns. On the one hand, mainstream hashtags tended to be more positive overall, with higher occurrences



2a. Top 10 mainstream hashtags

2b. Moderation evaders

Figure 2: Comparative emotion analysis of video descriptions

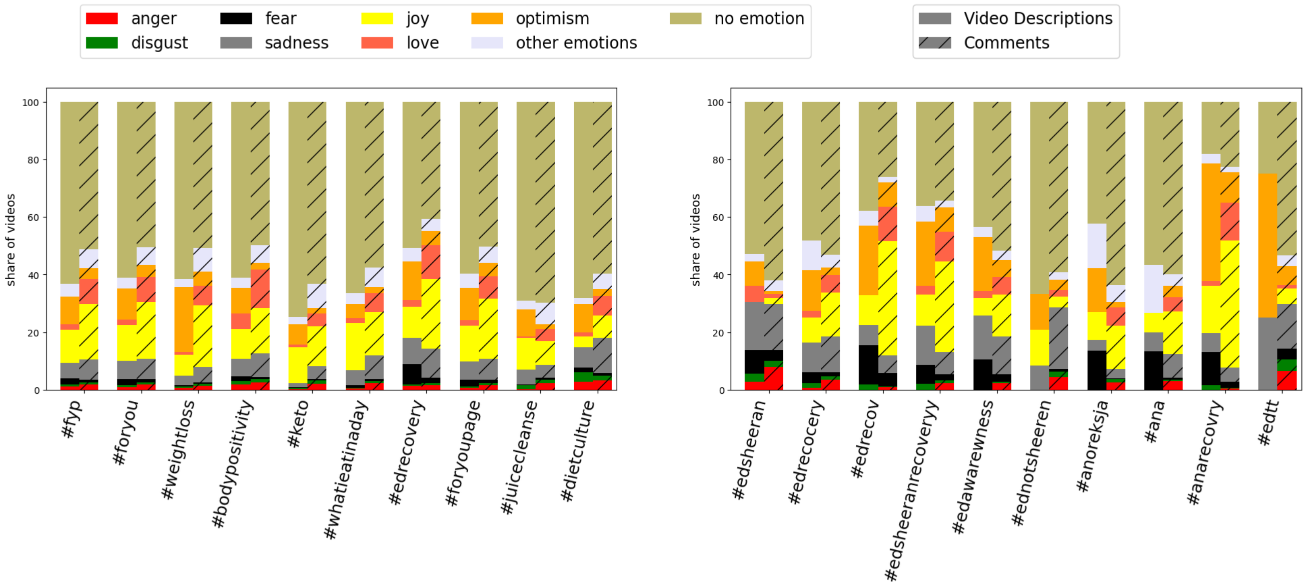
of joy and love. Moderation evaders displayed a more diverse emotional landscape, with notable expressions of negative sentiments such as fear and sadness, underscoring the potential risks associated with these hashtags. The identified emotional differences have implications for user exposure. Mainstream content, characterized by positive emotions, may contribute to a more uplifting user experience. On the other hand, the diverse emotional landscape of moderation evaders' content, including expressions of fear and sadness, suggests potential exposure to content with a more complex emotional impact, indicating both positive and potentially harmful messaging. Figure 5 lists sample video descriptions that express different emotions.

To better understand these differences, we conducted a manual review of the video content for posts containing moderation evaders. The videos containing the five moderation evaders with the highest number of occurrences (#edrecocery, #edsheeranrecoveryy, #edawarewness, #anarecvery, and #edrecov) often featured recovery content and positive messaging, providing tips on recovery approaches, highlighting the benefits of recovery, and addressing the toll of eating disorders.

Amongst the other five, the hashtags #anoreksja, #ednotsheeren, and #ana were not searchable, i.e., they are blocked when searching for them in the TikTok search bar, and links to mental health resources are provided instead (see Figure 4 in the Appendix). It should be noted that the mainstream hashtag #edrecovery was blocked as well.

Furthermore, an examination of posts containing #edtt revealed that many of the videos included displays of eating disorders-related behaviors, dark humor, and an overall glorification of eating disorders. Due to the co-occurrences of #edtt with more mainstream hashtags such as #edvent, #ed, and #fyp, this raises concerns about the potential exposure of users to harmful content of this nature. From the manual inspection of moderation evaders, we concluded they are not exclusively associated with harmful or negative content; instead, they encompass a variety of themes, including those focused on recovery and positive narratives related to eating disorders.

Findings and Remarks: In response to RQ2, we explored the emotional expressions in TikTok video descriptions, revealing differences between mainstream hashtags and moderation evaders. Mainstream content tends toward positivity, with joy and optimism prevalent, while moderation evaders' content shows a wider emotional range, including fear and sadness. A manual review of moderation evaders underscores the need to understand TikTok's emotional dynamics and associated risks across various hashtags while also showing that TikTok has increased its moderation efforts by blocking some of these hashtags.



3a. Top 10 mainstream hashtags

3b. Moderation evaders

Figure 3: Comparative emotion analysis of video descriptions and comments

RQ3: Emotions in Comments

Regarding RQ3, we investigated the relationship between emotional expressions in TikTok video descriptions and the emotions expressed in the user comments in response to these videos. Our analysis revealed that comments displayed a broader range of emotions compared to video descriptions themselves, with an increase in anger (Fig. 3).

In general, the emotional analysis of the video descriptions and comments of mainstream hashtags showed a consistent pattern. In posts tagged with recovery-related hashtags, both descriptions of videos and comments tended to be positive, with more expressions of joy, love, and optimism. This suggests that content associated with mainstream and recovery-related hashtags tends to foster positive engagement, creating a supportive and optimistic community atmosphere on TikTok. In comparison, the emotional analysis of video descriptions and comments associated with the moderation evaders reveals a clear emotional distinction. The user comments on these posts, while generally less emotionally charged than their corresponding video descriptions, still manifest a higher degree of joy, especially for the recovery-related hashtags.

Mainstream hashtags, while generally positive, exhibited slightly more anger in their comments. We also observed an increase in anger for the moderation evaders' posts comments in comparison to the video descriptions (see Fig. 3). The heightened expression of anger in comments, especially in moderation evaders' content, raises questions about the nature of discussions and interactions surrounding sensitive or controversial topics on TikTok. Figure 5 lists sample comments that express different emotions.

We have observed that some of the moderation evaders' comments displayed less emotion than the video descriptions, specifically #edsheeran, #edrecocery, #edawarewness #anoreskja, #ana, #anarecovry, and #edtt. Also, the comments for the hashtags #edsheeran, #ednotsheeren, and #edtt expressed less positive emotions compared to the video descriptions.

Interestingly, when analyzing the video descriptions for the hashtag #ednotsheeren, it is evident that positive emotions such as joy and optimism are present but relatively subdued, with percentages ranging from 0% to 12.5%. Notably, the emotion love is not expressed in the video descriptions. On the contrary, negative emotions, specifically sadness, stand out at 8.33%. Moving to comments, the disparity between positive and negative emotions becomes more pronounced. Positive emotions in comments, including joy, love, and optimism, collectively amount to 9.61%, significantly lower than the corresponding video description percentage, which was 25%. In contrast, negative emotions in comments, particularly sadness at 21.47%, surpass the negative emotions in the video descriptions. This discrepancy suggests that the content under #ednotsheeren may not be resonating positively with the audience, as indicated by the diminished expression of positive emotions in both video descriptions and comments.

Also - for the hashtag #edtt - even though the video descriptions contained mostly positive emotions (with 50% being labeled with optimism), this did not mean that the viewers exhibited the same sentiment. For the comments, the negative emotions total 29.86% while the positive emotions come to 12.98%. Despite 50% of the posts being classified as *optimism*, it appears that viewers might not perceive the content in the same way. This suggests that there may be a gap between the positive message in the video descriptions and how the audience actually interprets it.

Findings and Remarks: For RQ3, we examined the relationship between emotional expressions in video descriptions and comments. We observed that comments on videos with mainstream hashtags exhibited more emotions than video descriptions themselves, while moderation evaders generally elicited fewer emotions. Furthermore, there was a noticeable increase in negative emotions expressed in comments for moderation evaders compared to mainstream hashtags. This hints that some moderation evaders could hide potentially problematic content.

Discussion

We detected various TikTok communities related to eating disorders recovery, body positivity, and healthy living. These communities mostly contain content that is designed to be informational, inspirational, or uplifting in some shape or form; however, it should be noted that some of this well-intentioned content can have adverse effects as well. For example, content showing a person's body can encourage negative comparison and lead to viewers feeling poorly about their own bodies. Additionally, weight loss tips from unqualified TikTok users may be harmful to people struggling with eating disorders. Prior work done by the CCDH[1] and (Lerman et al. 2023) demonstrated that users within the ED community utilize misspellings, abbreviations, and the musical artist Ed Sheeran's name to avoid moderation. Within our dataset, we found 10 hashtags fitting this criterion, with each appearing in over 30 posts, all within the ED recovery/positivity community. These hashtags were found to cooccur with mainstream hashtags such as #fearfood, #mentalhealthmatters, and #recovery. This, combined with the fact that 1) the video descriptions with moderation evaders tend to evoke more negative emotions and 2) videos with some moderation evaders contain content that could promote ED behaviors, suggests that vulnerable users could potentially stumble upon harmful content even if they search for mainstream topics.

Nevertheless, the fact that some of these hashtags are blocked from the search page shows that TikTok has likely made an effort to moderate content that could potentially promote eating disorder behaviors, which may explain the overall positive tone of the content. The hashtag #edtt was one that was not blocked, though, and appeared to contain harmful content with "dark humor", ED-glorification, and ED-promoting advice. This is a hashtag that should be investigated further in future studies that analyze a larger amount of video content.

The emotional analysis for the video descriptions of the mainstream hashtags showed that they generally displayed a more balanced emotional distribution, representing a wide array of positive and neutral emotions. This could be because mainstream hashtags are often more visible to a broader audience which may lead users to adopt a more restrained and balanced emotional tone to appeal to a diverse viewership. On the other hand, moderation evaders' video descriptions tended to be more emotionally charged, with an increased presence of sadness, fear, and anger. This could suggest that those trying to avoid moderation are deliberately posting more negative content, believing that their content is less likely to be taken down.

The user comments on posts containing mainstream hashtags were more emotionally charged than their respective video descriptions. This could possibly be because users tend to be more careful and deliberate when sharing videos under popular hashtags, aiming to present a carefully curated image to a larger audience. These comments revealed a similar pattern to that of the video descriptions – joy and optimism were often the most common emotions. In contrast, the user comments on posts containing the moderation evaders were less emotionally charged than their respective video descriptions. Despite being less emotionally charged overall, these comments contained a higher amount of joy compared to their video descriptions, especially for hashtags related to recovery. This could indicate potential community formation where individuals show support, particularly concerning eating disorder recovery. Amongst both the comments for the mainstream hashtags and moderation evaders, anger was more present also - albeit still at a minimal level.

The discovery of moderation evaders, their co-occurrences with more mainstream hashtags, and the persistence of harmful content despite some blocked hashtags underscore the ongoing challenges in content moderation. The emotional analysis indicates that posts containing mainstream hashtags tend to display a balanced emotional distribution in the video description and comments, possibly due to broader visibility, whereas moderation evader hashtags exhibit more emotionally charged content, emphasizing the need for targeted moderation efforts to mitigate negative emotional impacts.

Ethics Statement

All data used for this study is public and collected following TikTok's terms of service. In our analysis of TikTok videos and comments within these videos, no identifiable information related to any user has been included and analysis was carried out on aggregated data. These steps ensure that negative outcomes due to use of these data are minimized.

The authors declare no competing interests.

Limitations

It is crucial to acknowledge the limitations of our analysis. The prevalence of "no emotion" in our analysis may be influenced by the hashtag-centric nature of TikTok video descriptions, potentially impacting the accuracy of emotional expression detection. Additionally, our emotional analysis was mostly confined to the textual video descriptions and comments, overlooking the visual and audio elements of the videos. While we manually reviewed some of the videos containing specific hashtags, this inspection was not comprehensive, and future research should look into analyzing these elements further while incorporating AI techniques and statistical analysis.

Conclusion

Our results provide valuable insights into the emotional dynamics of TikTok content concerning mental health and eating disorders – specifically content containing moderation evader hashtags. We have found that moderation evaders do co-occur with mainstream hashtags and that the video descriptions/comments of posts containing the former tend to be more emotionally charged (including a higher amount of negative emotions such as fear and sadness). The findings have implications for content creators, platform moderation, and interventions aimed at fostering a supportive online environment for discussions on mental health and eating disorders. Future studies should continue to inspect moderation evaders and add onto our findings by taking a comprehensive look at the video content of posts in addition to their video descriptions and comments.

References

Alhuzali, H.; and Ananiadou, S. 2021. SpanEmo: Casting multi-label emotion classification as span-prediction. *arXiv preprint arXiv:2101.10038*.

Branley, D. B.; and Covey, J. 2017. Pro-ana versus pro-recovery: A content analytic comparison of social media users' communication about eating disorders on Twitter and Tumblr. *Frontiers in psychology*, 8: 1356.

Casilli, A. A.; Pailler, F.; and Tubaro, P. 2013. Online networks of eating-disorder websites: why censoring pro-ana might be a bad idea. *Perspectives in public health*, 133(2): 1–2.

Chochlakis, G.; Mahajan, G.; Baruah, S.; Burghardt, K.; Lerman, K.; and Narayanan, S. 2023. Leveraging label correlations in a multi-label setting: A case study in emotion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Chua, T. H. H.; and Chang, L. 2016. Follow me and like my beautiful selfies: Singapore teenage girls' engagement in self-presentation and peer comparison on social media. *Computers in human behavior*, 55: 190–197.

Cobb, G. 2017. "This is not pro-ana": Denial and disguise in pro-anorexia online spaces. *Fat Studies*, 6(2): 189–205.

Cohen, R.; Newton-John, T.; and Slater, A. 2018. 'Selfie'-objectification: The role of selfies in self-objectification and disordered eating in young women. *Computers in Human Behavior*, 79: 68–74.

Cooper, M.; Reilly, E. E.; Siegel, J. A.; Coniglio, K.; Sadeh-Sharvit, S.; Pisetsky, E. M.; and Anderson, L. M. 2022. Eating disorders during the COVID-19 pandemic and quarantine: an overview of risks and recommendations for treatment and early intervention. *Eating disorders*, 30(1): 54–76.

Corzine, A.; and Harrison, V. 2023. Social Contagion, from Suicide to Online Challenges to Eating Disorders: Current Research and Harm Mitigation Strategies for Youth Online. *Journal of Online Trust and Safety*, 2(1).

de Valle, M. K.; and Wade, T. D. 2022. Targeting the link between social media and eating disorder risk: A randomized controlled pilot study. *International Journal of Eating Disorders*, 55(8): 1066–1078.

Dondzilo, L.; Rodgers, R. F.; and Dietel, F. A. 2023. Association between engagement with appearance and eating related TikTok content and eating disorder symptoms via recommended content and appearance comparisons. *International Journal of Eating Disorders*.

Fardouly, J.; Diedrichs, P. C.; Vartanian, L. R.; and Halliwell, E. 2015. Social comparisons on social media: The impact of Facebook on young women's body image concerns and mood. *Body image*, 13: 38–45.

Festinger, L. 1957. Social comparison theory. *Selective Exposure Theory*, 16: 401.

Fredrickson, B. L.; and Roberts, T.-A. 1997. Objectification theory: Toward understanding women's lived experiences and mental health risks. *Psychology of women quarterly*, 21(2): 173–206.

Frieiro Padin, P.; González Rodríguez, R.; Verde Diego, M. D. C.; Vázquez Pérez, R.; et al. 2021. Social media and eating disorder psychopathology: A systematic review. *Cyberpsychology Journal of Psychosocial Research on Cyberspace*.

Greene, A. K.; and Norling, H. N. 2023. "Follow to* actually* heal binge eating": A mixed methods textual content analysis of# BEDrecovery on TikTok. *Eating Behaviors*, 50: 101793.

Greene, A. K.; Norling, H. N.; Brownstone, L. M.; Maloul, E. K.; Roe, C.; and Moody, S. 2023. Visions of recovery: a cross-diagnostic examination of eating disorder pro-recovery communities on TikTok. *Journal of Eating Disorders*, 11(1): 109.

Hawkins, N.; Richards, P. S.; Granley, H. M.; and Stein, D. M. 2004. The impact of exposure to the thin-ideal media image on women. *Eating disorders*, 12(1): 35–50.

Herrick, S. S.; Hallward, L.; and Duncan, L. R. 2021. "This is just how I cope": An inductive thematic analysis of eating disorder recovery content created and shared on TikTok using# EDrecovery. *International journal of eating disorders*, 54(4): 516–526.

Hogue, J. V.; and Mills, J. S. 2019. The effects of active social media engagement with peers on body image in young women. *Body image*, 28: 1–5.

Holland, G.; and Tiggemann, M. 2016. A systematic review of the impact of the use of social networking sites on body image and disordered eating outcomes. *Body image*, 17: 100–110.

Hülsing, G. 2021. *# Triggerwarning: Body Image: A qualitative study on the influences of TikTok consumption on the Body Image of adolescents*. B.S. thesis, University of Twente.

Hung, M.; et al. 2022. A content analysis on fitspiration and thinspiration posts on TikTok. *Cornell Undergraduate Research Journal*, 1(1): 55–62.

Iqbal, M. 2021. TikTok revenue and usage statistics (2021). *Business of apps*, 1(1).

Jiotsa, B.; Naccache, B.; Duval, M.; Rocher, B.; and Grall-Bronnec, M. 2021. Social media use and body image disorders: Association between frequency of comparing one's own physical appearance to that of people being followed on social media and body dissatisfaction and drive for thinness.

International journal of environmental research and public health, 18(6): 2880.

Karsay, K.; Trekels, J.; Eggermont, S.; and Vandenbosch, L. 2021. "I (don't) respect my body": Investigating the role of mass media use and self-objectification on adolescents' positive body image in a cross-national study. *Mass Communication and Society*, 24(1): 57–84.

Keski-Rahkonen, A. 2021. Epidemiology of binge eating disorder: prevalence, course, comorbidity, and risk factors. *Current opinion in psychiatry*, 34(6): 525–531.

Kleemans, M.; Daalmans, S.; Carbaat, I.; and Anschütz, D. 2018. Picture perfect: The direct effect of manipulated Instagram photos on body image in adolescent girls. *Media Psychology*, 21(1): 93–110.

Kruglanski, A. W.; Gelfand, M. J.; Bélanger, J. J.; Sheveland, A.; Hetiarachchi, M.; and Gunaratna, R. 2014. The psychology of radicalization and deradicalization: How significance quest impacts violent extremism. *Political Psychology*, 35: 69–93.

Lerman, K.; Karnati, A.; Zhou, S.; Chen, S.; Kumar, S.; He, Z.; Yau, J.; and Horn, A. 2023. Radicalized by Thinness: Using a Model of Radicalization to Understand Pro-Anorexia Communities on Twitter. *arXiv preprint arXiv:2305.11316*.

Liu, J. 2022. Social Media and Its Impact on Chinese's Women Body Image: The Effects of Body Comparison and Motivation for Social Media use. In *2021 International Conference on Public Art and Human Development (ICPAHD 2021)*, 206–210. Atlantis Press.

Lonergan, A. R.; Bussey, K.; Fardouly, J.; Griffiths, S.; Murray, S. B.; Hay, P.; Mond, J.; Trompeter, N.; and Mitchison, D. 2020. Protect me from my selfie: Examining the association between photo-based social media behaviors and self-reported eating disorders in adolescence. *International Journal of Eating Disorders*, 53(5): 755–766.

Manaf, N. A.; Saravanan, C.; and Zuhrah, B. 2016. The prevalence and inter-relationship of negative body image perception, depression and susceptibility to eating disorders among female medical undergraduate students. *Journal of clinical and diagnostic research: JCDR*, 10(3): VC01.

Marks, R. J.; De Foe, A.; and Collett, J. 2020. The pursuit of wellness: Social media, body image and eating disorders. *Children and youth services review*, 119: 105659.

McCashin, D.; and Murphy, C. M. 2023. Using TikTok for public and youth mental health–A systematic review and content analysis. *Clinical Child Psychology and Psychiatry*, 28(1): 279–306.

Mink, D. B.; and Szymanski, D. M. 2022. TikTok use and body dissatisfaction: Examining direct, indirect, and moderated relations. *Body Image*, 43: 205–216.

Peng, S. 2023. The Negative Relationship Between Social Media and Body Image of Women. *Journal of Education, Humanities and Social Sciences*, 22: 557–562.

Pruccoli, J.; De Rosa, M.; Chiasso, L.; Perrone, A.; and Parmeggiani, A. 2022. The use of TikTok among children and adolescents with Eating Disorders: Experience in a third-level public Italian center during the SARS-CoV-2 pandemic. *Italian Journal of Pediatrics*, 48(1): 138.

Rando-Cueto, D.; de las Heras-Pedrosa, C.; and Paniagua-Rojano, F. J. 2023. Health Communication Strategies via TikTok for the Prevention of Eating Disorders. *Systems*, 11(6): 274.

Rodgers, R. F.; Lombardo, C.; Cerolini, S.; Franko, D. L.; Omori, M.; Fuller-Tyszkiewicz, M.; Linardon, J.; Courtet, P.; and Guillaume, S. 2020. The impact of the COVID-19 pandemic on eating disorder risk and symptoms. *International Journal of Eating Disorders*, 53(7): 1166–1170.

Rodgers, R. F.; and Rousseau, A. 2022. Social media and body image: Modulating effects of social identities and user characteristics. *Body Image*, 41: 284–291.

Salomon, I. P. 2020. Examining the Impact of Social Media Use on Body Dissatisfaction and Eating Disorder Symptomatology Among Adolescents.

Schlegl, S.; Maier, J.; Meule, A.; and Voderholzer, U. 2020. Eating disorders in times of the COVID-19 pandemic—Results from an online survey of patients with anorexia nervosa. *International Journal of Eating Disorders*, 53(11): 1791–1800.

Sha, P.; and Dong, X. 2021. Research on adolescents regarding the indirect effect of depression, anxiety, and stress between TikTok use disorder and memory loss. *International journal of environmental research and public health*, 18(16): 8820.

Tiggemann, M.; Anderberg, I.; and Brown, Z. 2020. Uploading your best self: Selfie editing and body dissatisfaction. *Body Image*, 33: 175–182.

Walsh, B. T.; Attia, E.; and Glasofer, D. R. 2020. *Eating disorders: What everyone needs to Know®*. Oxford University Press.

Wang, E. L.; Luceri, L.; Pierri, F.; and Ferrara, E. 2023. Identifying and characterizing behavioral classes of radicalization within the qanon conspiracy on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 890–901.

Westenberg, J. M.; and Oberle, C. D. 2023. The impact of body-positivity and body-checking TikTok videos on body image. *The Journal of Social Media in Society*, 12(1): 49–60.

Wick, M. R.; and Keel, P. K. 2020. Posting edited photos of the self: Increasing eating disorder risk or harmless behavior? *International Journal of Eating Disorders*, 53(6): 864–872.

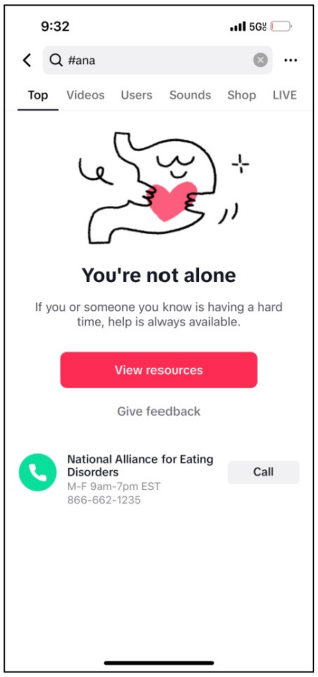
Zhang, J.; Wang, Y.; Li, Q.; and Wu, C. 2021. The relationship between SNS usage and disordered eating behaviors: a meta-analysis. *Frontiers in Psychology*, 12: 641919.

Appendix

Search Terms thinspo, proana, proanatips, anatips, meanspo, fearfood, sweetspo, eatingdisorder, bonespo, promia, redbracetpro, m34nspo, fatspo, lowcalrestriction, edvent, WhatIEatInADay, Iwillbeskinny, thinspoa, ketodiet, skinnycheck, thighgapworkout, bodyimage, bodygoals, weightloss, skinnydiet, chloetingchallange, fatacceptance, midriff, foodistheenemy, cleanvegan, keto, cleaneating, intermittentfasting, juicecleanse, watercleanse, EDrecovery, bodypositivity, dietculture.

Table 2: Co-occurrences of Moderation Evader Hashtags with Some Mainstream Hashtags

Hashtag	Total Occurrences	#recovery	#ed	#edvent
#edrecocery	138	69	64	21
#edsheeranrecoveryy	102	57	31	16
#edawarewness	86	43	38	37
#anarecovry	61	55	< 10	< 10
#edrecov	58	44	22	16
#anoreksja	47	38	37	< 10
#edtt	39	< 10	35	39
#edsheeran	38	15	< 10	14
#ednotsheeren	36	< 10	22	31
#ana	31	16	25	< 10



You're not alone

If you or someone you know is having a hard time, help is always available.

National Alliance for Eating Disorders
M-F 9am-7pm EST
866-662-1235

Figure 4: Hashtag #ana blocked on TikTok, redirecting users to mental health resources

`Emotion:` Joy
`Video Description:` I did it! Let's get the big girls to go viral #fy #fat #big #bbw #curvy #plussize #biggirl #bodypositivity #fatacceptance #stitch #ArmaniMyWay
`Comment:` Think about how sweet and amazing it tastes remember the feeling of accomplishment once you have finished eating xx Stay strong

`Emotion:` Optimism
`Video Description:` so much work to do but we gon' do it! #dietculture #bodyimage #caloriedeficit #selflove #toxicrelationship #edrecovery
`Comment:` Im going to watch this everyday until i reach my goal this is real motivations right here

`Emotion:` Sadness
`Video Description:` Spent a long time wishing things like this would happen, just to shrink my body :(#PlusSize #FatLiberation #FatAcceptance #BodyPositivity #FatTikTok #ThatFatBaddee #fatacceptance #stitch #ArmaniMyWay
`Comment:` I started crying when it said "Go help other girls!"

`Emotion:` Fear
`Video Description:` Not knowing the exact amount was very scary, but I'm so sick of this scale!! #ed #recovery #eatittobeatit #edrecovery #edrecocery #foodisfuel #fearfood #fearfoodchallenge #fyp #fy
`Comment:` As a nutritionist this is TERRIFYING. Whyyyyy would they even publish this?

Figure 5: Samples of video descriptions and comments expressing emotions