Hashtag-based cluster analysis of eating disorder recovery content on TikTok with #edrec0very

Ammar Alabboodi
ana154@pitt.edu

Havannah Tung
hat127@pitt.edu

ABSTRACT

Since TikTok started to gain its popularity among young demographics in early 202, many research studies have shown concerns that the social media platform lacks regulations against content that promotes unrealistic beauty ideals. As a result, in September 2020, TikTok improved its ad policies on weightl loss and dieting products. In February 2021, the social media platform added the feature to redirect searches related to eating disorders (ED) to the National Alliance for Eating Disorder Helpline. The hashtag #edrecovery was one of those searches. Misspelled hashtags such as "#edrec0very" started to occur as a workaround. The goal for this project is to use unsupervised learning to conduct hashtag analysis to understand common themes of TikToks with the workaround hashtag.

KEYWORDS

Data mining, clustering, mental health, social media, eating disorders, TikTok, unsupervised learning

1. INTRODUCTION

This project is inspired by the thematic analysis conducted by Herrick et. al [1] on hashtag #edrecovery on TikTok, which was published in December 2020. In February 2021, TikTok started to mute eating-disorders-related searches. When an ED related keyword is searched, a page titled "You're not alone - If you or someone you know is having a hard time, help is always available", followed by the helpline to reach the National Alliance for Eating Disorders, would show up as the search result. Surprisingly, the seemingly pro-recovery hashtag #edrecovery is among these muted searchyes.

Intentionally misspelled hashtags started to emerge as workarounds after the regulation. For example, #edrec0very, where the "o" is replaced with "zero", has become of the popular work-around hashtags for the pro-recovery community on TikTok.

The goal for this project is to use clustering to find patterns in ED-related TikTok to answer three research questions: :

1. What are the common themes posted with the hashtag #edrec0very?
2. Are certain themes gain more popularity than others?
3. For creators who have posted ED-recovery-related TikTok, what type of content are they generally posting? Are they ED-specific?

2. RELATED WORK

Most research that studies ED-recovery content on social media is conducted by qualitative analyses. For example, Goh et al. conduct qualitative content analysis on Instagram posts with #EatingDisordersRecovery and #EDRecovery, where 405 posts are reviewed by researchers and qualitatively coded based on a codebook. Similarly, Hedrick et al. and Greene et al. use similar framework to understand the themes on #edrecovery (n=150) and diagnostic-specific ED-recovery hashtags (n=241), respectively, on TikTok.

One of the advantages of qualitative analysis is its ability to help gain deeper insights on ED-recovery online community. However, its biggest drawback is its time and labor consuming nature.

Quantitative research on ED-recovery content on TikTok is still sparse. Following their qualitative thematic analysis on diagnosis-specific recovery TikTok, Greene et al. experiment using both qualitative and quantitative analyses to understand #BEDrecovery (binge eating disorder recovery). The researchers determine a list of keywords related to the topic, and conduct textual analysis 10xx videos.

With the large amount of social media data being generated daily, data mining techniques will help us understand large dataset more timely. Our project aims to use clustering techniques to generate preliminary results to understand common themes in #edrec0very, which is experimental yet innovative given that research using machine learning to study the subject is still sparse. Despite that it would be hard to evaluate our method by comparing to similar research work, we hope that with implementing qualitative

analysis in the future, we will be able to better assess the quality of our project.

3. DATASET

3.1. Search results of #edrec0very

We use Apify TikTok Data Extractor to collect the #edrec0very search result data. The extractor scrapes TikTok videos and their creators' metadata based on search query. We enter "edrec0very" as input and set the number of videos to be scraped to unlimited to maximize the videos we get. The data is extracted on March 31,2025.

The raw dataset consists of 127 videos with features including several author metadata, video metadata, audio metadata, like, share, and view counts, and hashtag used for each video, etc.

After cleaning the data, the dataset includes 10 features: *authorMeta.digg, authorMeta.fans, authorMeta.following, authorMeta.heart, authorMeta.video, collectCount, diggCount, playCount, shareCount,* and *hashtags.* The numerical data is then scaled and normalized before analysis (Figure 1).

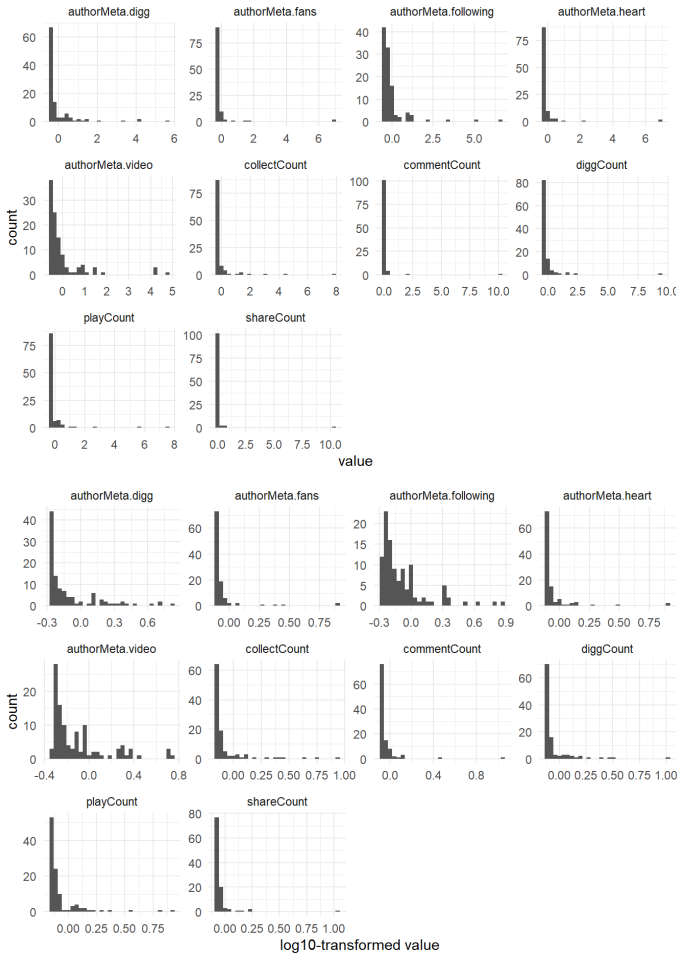


Figure 1. Distribution of the ten numerical features before and after log10 transformation

We also generate the top 20 hashtags used in the videos that show up in the search result (Figure 2). The top 20 hashtags are: #edrec0very, #fyp, #mentalhealth, #recovery, #recoveryispossible, #ed, #edsheeranrecoveryy, #mentalhealthmatters, #foryou, #anarec0very, #foryoupage, #neda, #relatable, #3drecovery, #ana, #edrecovry, #recoverytok, #ednotsheeranrec0very, #bed, and #edawareness.

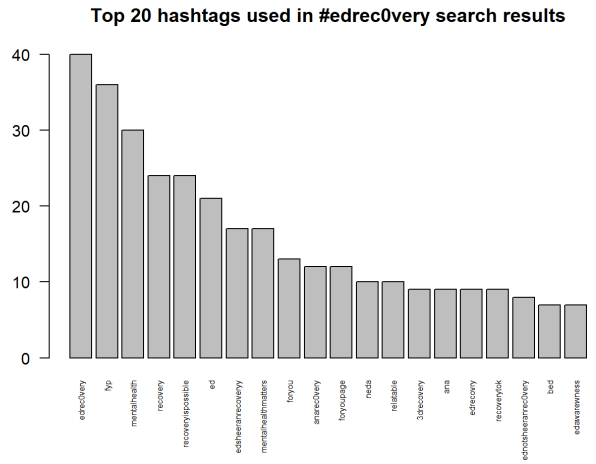


Figure 2. Top 20 hashtags used in content showing up in #edrec0very search result

3.2. Hashtag data of profiles shown up in #edrec0very search results

We generate the list of usernames from the search result dataset and modify the Apify TikTok Data Extractor using Python to capture up to 50 videos and all the hashtags used for each creator.

4771 videos and their hashtags are captured. We delete videos that use the same combination of hashtags if they are posted by the same creator, which leads to 2565 videos left in the dataset.

Similarly, we generate the top 20 hashtags co-occurring with #edrec0very in 106 user profiles (Figure 3).

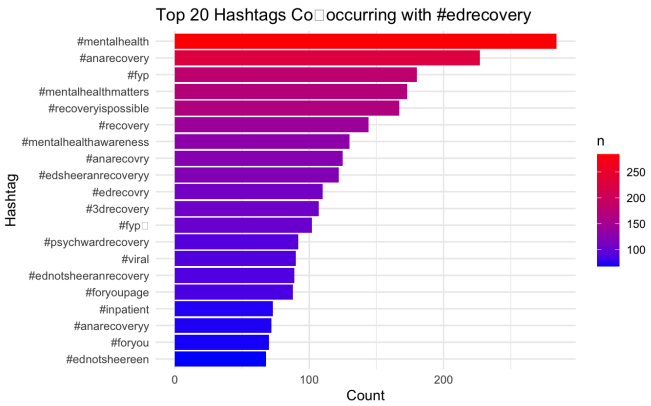


Figure 3. Top 20 hashtags used by the 106 content creators whose TikToks showed up in #edrec0very search result

4. METHOD

4.1 Method for #edrec0very search result analysis

We first turn all hashtags into binary dummy variables to conduct Principal Component Analysis (PCA) on all hashtag variables. The PCA results show that the first principal components (PC) are sufficient for clustering.

We arbitrarily pick four clusters (k = 4) prior to k-means clustering. After clustering, the statistics of numerical variables indicating the popularity of the videos are calculated for each cluster for further analysis and evaluation for clustering result.

4.2 Method for user profile hashtag analysis

The preprocessing pipeline prior to clustering for user profile hashtag analysis includes:

1. Parsing and normalizing hashtags (lowercasing).
2. Building a **Document-Term Matrix (DTM)** from the list of hashtags associated with each video.
3. Applying **TF-IDF weighting** to emphasize informative, discriminative hashtags.
4. Conducting **k-means clustering** on the TF-IDF representations to identify thematic groupings.
5. Reducing the high-dimensional TF-IDF space using **Principal Component Analysis (PCA)** for visualization.

Finally, we interpreted clusters by examining the top co-occurring hashtags and common content themes within each cluster to evaluate our clustering result.

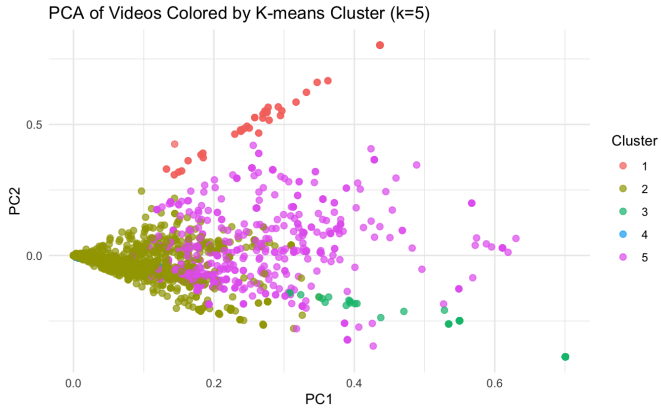


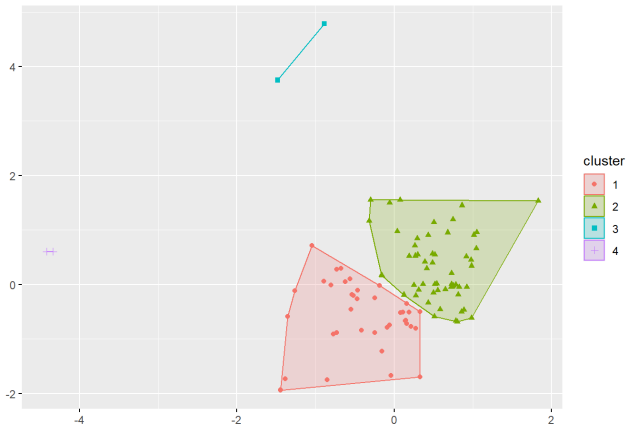
Figure 4. PCA of videos colored based on hashtag-based k-means clustering (k = 5) result for user profile dataset

5. EVALUATION RESULT

5.1 Clustering results for #edrec0very search results and evaluation

The top hashtags used by videos with each cluster are as followed and in Figure 5:

Cluster 1 (n = 53): #edrec0very (n = 33), #mentalhealth (n = 17), #fyp (n = 14), #recoveryispossible (n = 14), #mentalhealthmatters (n = 13), #edsheeranrecoveryy (n = 12).
Cluster 2 (n = 70): #fyp (n = 17), #ed (n = 16), #recovery (n = 16), #mentalhealth (n = 10), #recoveryispossible (n = 10).
Cluster 3 (n = 2): #gymtok (n = 2), #binge (n =2), #bingefree (n = 2), #caloriedeficit (n = 2)
Cluster 4 (n = 3): #3drecovery (n = 3), #anarec0very (n = 3),



#anarecovry (n = 3), #edrec0very (n = 3).
Figure 5. Hashtag-based K-means clustering result (k = 4) for #edrec0very search result

The size of the clusters is not quite even. Cluster 1 and 2 have significantly more members than Cluster 3 and 4. However, by comparing the top hashtags for Cluster 1 and 2, the two clusters share high thematic similarities On the other hand, even though Cluster 3 and 4 have very low number of members, both clusters seemingly have distinguish themes compared to the other groups. The top hashtags of Cluster 3 suggest diagnosis-specific content with beinge eating disorder, and those of Cluster 4 suggest the use of intentionally misspelled workaround hashtags for the muted #edrecovery.
The popularity statistics for each cluster (Figure 6) suggest the following:
1. Videos with diagnosis-specific hashtags, in contrast to general mental health or ED-related hashtags,
2. Even though creators posted content with workaround hashtags seems to be more popular than the other creators, their video is still yet as popular as the others, suggesting that the workaround hashtags are still gaining popularity.

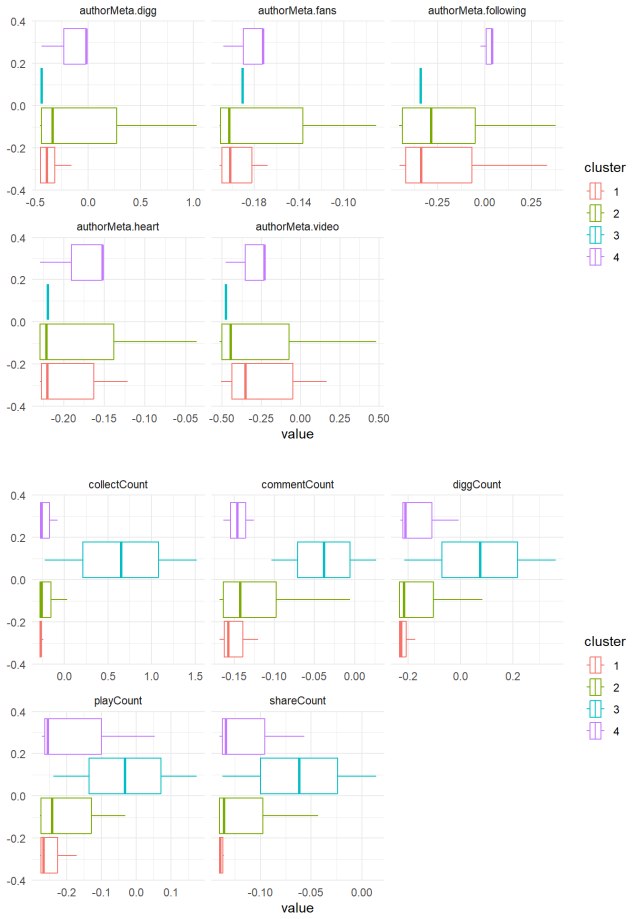


Figure 6. Popularity statistics for each hashtag-based cluster

5.2 Clustering results for user profile dataset and evaluation

We evaluated the clustering results using top hashtag proportion and entropy as indicators of cluster coherence and diversity (Figure 7). Cluster 3, centered around the hashtag #exerciseaddictionrecovery, showed the highest thematic consistency, with a top tag proportion of 77.8% and a low entropy of 0.76. This indicates a tightly focused group where the majority of posts relate directly to recovery from exercise addiction. Similarly, Cluster 4, associated with #fitcheck, also exhibited high coherence, with 76.9% of posts containing the top tag and a relatively low entropy of 0.99, suggesting a clear thematic grouping around fitness-related self-expression. Cluster 1, dominated by #collegelife, showed moderate coherence, with a top tag proportion of 44.4% and an entropy of 1.53. This suggests that while college-related experiences form a substantial theme, there is still a noticeable degree of hashtag diversity

within the cluster. In contrast, Cluster 2, where #1 was the top tag representing only 9.1% of posts, had a higher entropy of 5.56, indicating a much more heterogeneous group with a wide variety of themes and little concentration around a single topic. Cluster 5, stood out as the most diverse and least cohesive group: although #1 was again the top hashtag, it accounted for only 5.1% of the posts, and the entropy was extremely high at 8.73. This suggests that Cluster 5 is highly mixed, with no clear dominant theme and a wide range of unrelated content. Overall, the evaluation shows that some clusters captured well-defined themes related to specific aspects of recovery and fitness, while others were much more diffuse, reflecting the broad and varied nature of social media discussions in this domain.

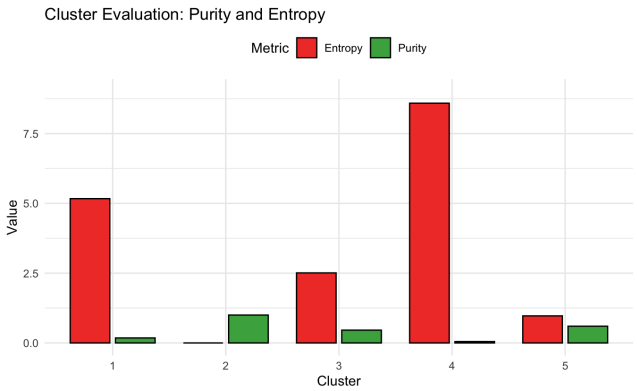


Figure 7. Purity and entropy for each cluster

6. Discussion

6.1 Limitations

While our project produced meaningful insights into eating disorder recovery content on TikTok, several limitations must be acknowledged. First, there was an inherent sampling bias due to the use of Apify scraping, which favored popular and recent posts. As a result, the dataset may not represent the full diversity of content creators or recovery narratives on the platform. Furthermore, although we collected hundreds of observations, the dataset remains relatively small when considering the scale and rapid growth of TikTok. The dynamic nature of the platform, where trends and popular hashtags shift daily, also poses challenges for the stability and reproducibility of our findings. Another limitation stems from the technical constraints of scraping itself; Apify scraping cannot guarantee access to all videos matching a search, and factors like platform updates, account privacy settings, and API rate limits can introduce omissions in the data. Lastly, there was no available ground truth for thematic labeling, making cluster evaluation primarily unsupervised and dependent on interpretation of hashtag patterns rather than validated categories.

6.2 Future directions

Looking ahead, there are several promising directions for future work. One key improvement would be obtaining larger and more representative datasets, potentially by leveraging official TikTok APIs or collaborating with research teams that have privileged access to richer data. Expanding the analysis to include a wider variety of related "turn-around" hashtags, beyond #edrecovery and #edrec0very, could uncover broader community patterns and hidden subcultures. In addition, incorporating dynamic clustering methods to analyze how user communities and thematic trends evolve over time would yield valuable longitudinal insights. Another potential extension would be the integration of other user metadata, such as follower counts or account longevity, into the clustering features to better characterize different creator types. Furthermore, human-based content analysis could be used alongside computational methods to validate and deepen the interpretation of cluster themes. Finally, exploring more sophisticated representations, such as using neural embeddings (e.g., BERT or sentence-transformer models for captions and descriptions), may capture richer semantic relationships between posts than traditional TF-IDF approaches allow.

7. Conclusions

This project explored eating disorder recovery content on TikTok by analyzing hashtag usage patterns around #edrec0very and related hashtags. Using clustering and dimensionality reduction, we identified distinct thematic groups among both individual videos and broader user posting behaviors.

Our findings highlight both the supportive and concerning aspects of ED-related discourse on TikTok, including authentic recovery narratives as well as lingering pro-ED themes. Despite limitations related to data collection and sample size, the project demonstrates the potential of unsupervised text mining techniques for understanding online health communities.

In future work, larger and more diverse datasets, combined with deeper content analysis, will further enhance our understanding of how young people engage with eating disorder recovery content on social media platforms like TikTok.

ACKNOWLEDGMENTS

This project serves as the final project for Dr. Yu-ru Lin's Data Mining course at the University of Pittsburgh. Spring 2025.

REFERENCES

[1] Herrick SSC, Hallward L, Duncan LR. "This is just how I cope": An inductive thematic analysis of eating disorder recovery content created and shared on TikTok using #EDrecovery. *Int J Eat Disord*. 2021;54(4):516-526. doi:10.1002/eat.23463

[2] Greene AK, Norling HN, Brownstone LM, Maloul EK, Roe C, Moody S. Visions of recovery: a cross-diagnostic examination of eating disorder pro-recovery communities on TikTok. *J Eat Disord*. 2023;11(1):109. Published 2023 Jul 3. doi:10.1186/s40337-023-00827-7

[3] Au ES, Cosh SM. Social media and eating disorder recovery: An exploration of Instagram recovery community users and their reasons for engagement. *Eat Behav*. 2022;46:101651. doi:10.1016/j.eatbeh.2022.101651

[4] Goh AQY, Lo NYW, Davis C, Chew ECS. #EatingDisorderRecovery: a qualitative content analysis of eating disorder recovery-related posts on Instagram. *Eat Weight Disord*. 2022;27(4):1535-1545. doi:10.1007/s40519-021-01279-1

[5]. Greene AK, Norling HN. "Follow to *actually* heal binge eating": A mixed methods textual content analysis of #BEDrecovery on TikTok. Eat Behav. 2023 Aug;50:101793. doi: 10.1016/j.eatbeh.2023.101793. Epub 2023 Aug 24.