EDA Report of TikTok Datasets

Weilu Sun

02.02.2024

1. Introduction

This Exploratory Data Analysis is based on two datasets, which are “Urls_Scrape” and “comments”.

The “Urls_Scrape” dataset, which stores information based on the URL of each video, includes data about the user who posted the video (user ID, user name, whether the user is verified), the basic information about the video (the description the user wrote, challenges the user mentioned, people tagged by the user), music information of the video (title, author of the music), technical parameters of the video (id, ratio, height, width, create time, duration, bitrate, encoded type, format, quality, codec type, definition), other users’ interaction with this video (the number of likes, share, comment, and play of the video).

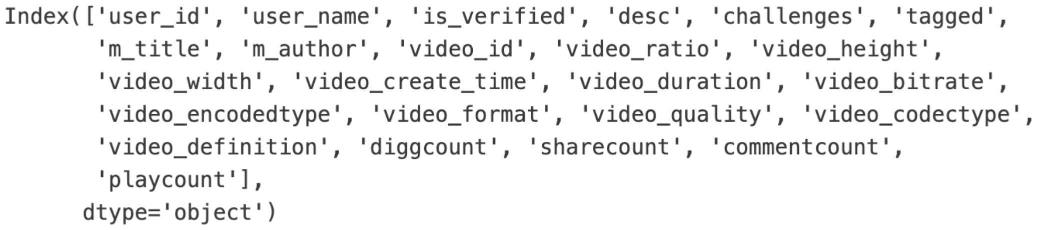
The “comments” dataset, which stores the comments information of each video, includes data of user ID (who commented on the video), text (the content of the comment), number of likes (of the comment), and the user name (of who commented on the video).

The main objective of this EDA is to explore the dataset to find out if they have some features potential for further study, or if there’re any relationships between variables valuable to know, and try to get some clues for interesting questions.

2. Data Cleaning, Visualization, and Analysis

2.1 “Url” data

First, I converted the dictionary-formatted data into a data frame. Then I extract keywords we are interested in, including user ID, description of the video, number of likes the video got, etc, from the dictionaries into new columns of the data frame. Then I drop the previous general [‘user,’ ‘text,’ ‘music,’ ‘video,’ ‘stats’] columns. Therefore, I got a data frame consisting of url-related data with 14816 rows and 24 columns. Columns’ names are shown in Figure 1.



```
Index(['user_id', 'user_name', 'is_verified', 'desc', 'challenges', 'tagged',
       'm_title', 'm_author', 'video_id', 'video_ratio', 'video_height',
       'video_width', 'video_create_time', 'video_duration', 'video_bitrate',
       'video_encodedtype', 'video_format', 'video_quality', 'video_codectype',
       'video_definition', 'diggcount', 'sharecount', 'commentcount',
       'playcount'],
      dtype='object')
```

Figure 1. All Columns in “Url” Data Frame

I found that there were 143 missing data in “music author” and 222 missing data in “video bitrate,” “video encoded type,” “video format,” “video quality,” “video codec type,” and “video definition”, which were shown in Figure 2. The yellow lines there represent the missing data. Compared to the total number of data we have, which is 14816, and these variables might not be the main features I will dive deep into, I think it’s fine to leave the missing data there.

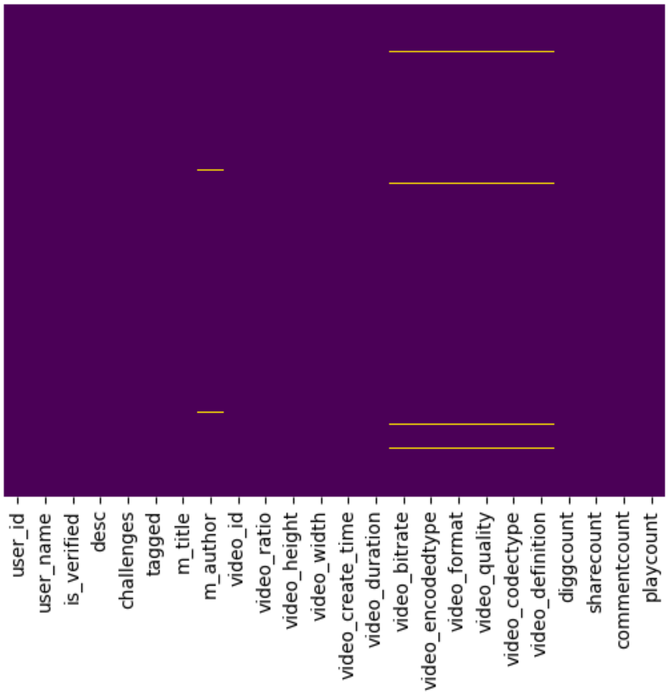


Figure 2. Missing Data in "Url" Dataset

I explored how many users posting their videos in this sample are verified and how many are not by creating the count plot shown in Figure 3. From this figure, we can see that from this sample, most video posters are not verified. While the verified accounts show that they're confirmed by TikTok to belong to the person or brand they represent, many times, unverified accounts represent "normal" users.

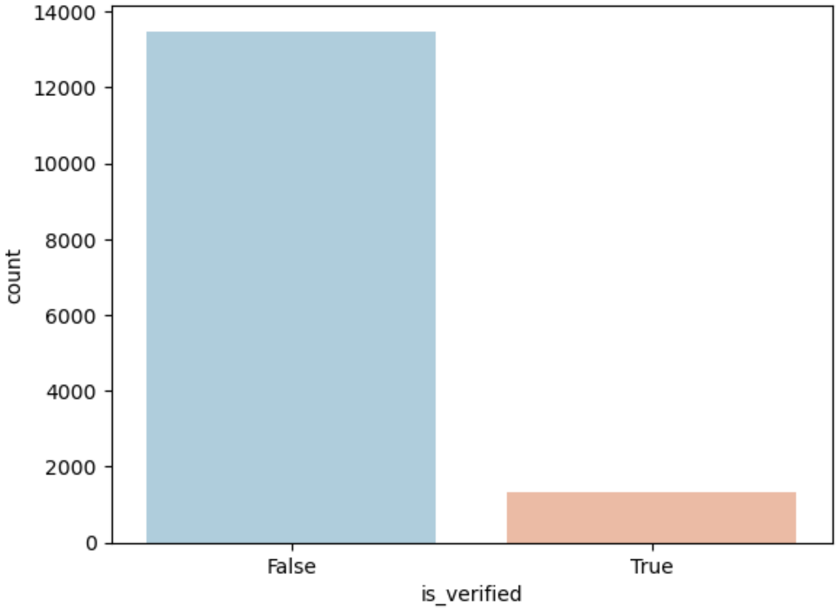


Figure 3. Number of (Un-)Verified Users

Next, I took a look at the most used "hashtags" among these videos. I got the top 20 hashtags people like to include in their descriptions of the videos shown below. The top two are hashtags "fyp" and "foryou", which are mostly used for being recommended to other users' default recommended pages. Afterward, I found some more interesting hashtags including "weightloss", "bodypositivity", "keto", "whatieatinaday", "edrecovery", "juicecleanse", "bodyimage", "intermittentfasting", "dietculture", "fearfood", "fatacceptance", "cleaneating", "bodygoals", "food", "fitness", and "recovery". From these hot hashtags, I find that people posting these videos care about weight loss, body goals while at the same time advocating body positivity. They're also trying to hit the goal through quote-unquote healthy methods, which are popular in recent days, including "juice cleanse", "intermittent fasting", "clean eating" an so on. But there are also people struggling with some problems associated with unbalanced eating and exercising show, focusing on how to beat the fear of food, how to recover from eating disorders, how to work better with diet culture, etc. The hashtags indicate that videos in this sample are mostly about healthy eating habits and balanced lifestyles. Most hashtags tell us people are trying to adopt "healthy" ways to lose weight and stay healthy.

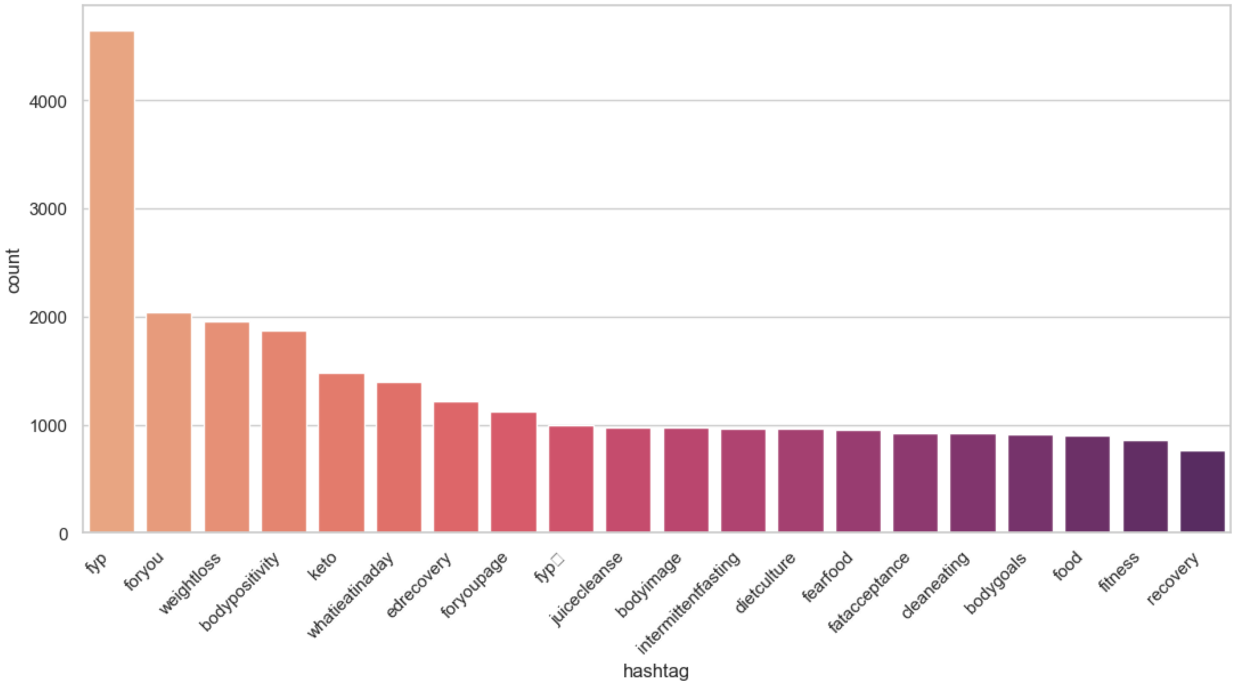


Figure 4. Top 20 Hashtags from All Videos

I also created a distribution plot to have an idea of the distribution of hashtags.

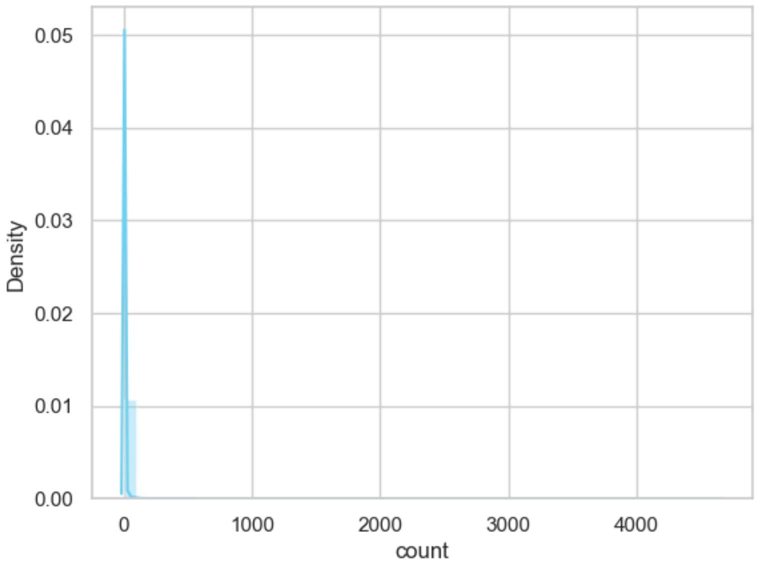


Figure 5. Distribution of Hashtags

It’s clear for us to see that most time, the hashtags will only show up once. There’s a low probability that the shown times of a hashtag will be great. That means most hashtags shown in the dataset will only have a few times of appearance, while few hashtags have a very high frequency.

Besides “hashtags”, I also list out the most frequently mentioned “challenges” in a similar way. The results are shown as follows. The hot “challenge” keywords are pretty similar to the “hashtag” ones.

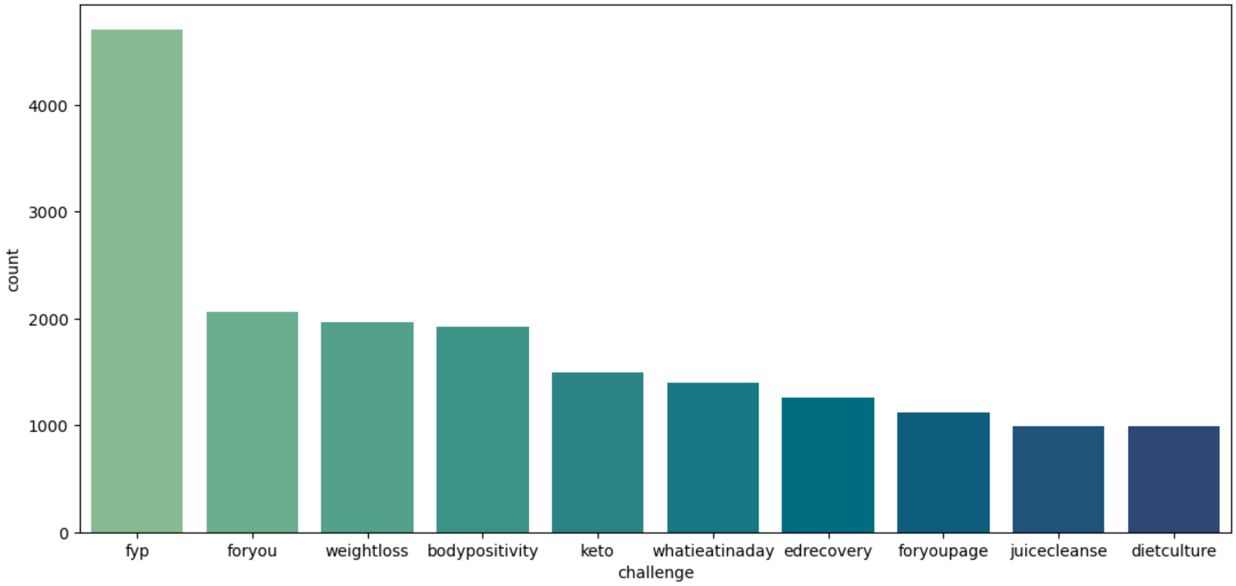


Figure 6. Frequently Mentioned Challenges

Next, I explored the volume of videos across different time periods. I found that there are 5857 videos in 2022, 4363 videos in 2021, 2332 videos in 2020, 104 videos in 2019, and 8 videos before 2019. The earliest video in this sample was created on December 30, 2016, while the latest video was created on April 20, 2023. Considering that TikTok was founded in 2016 and became popular in around 2020 and 2021, it matched the increase in the number of videos in this sample.

Table 1. Number of Videos across Years

Year	Number of Videos
2022	5857
2021	4363
2020	2332
2019	104
Before 2019	8

I also drew a distribution plot to have a clearer look on the increase of videos in TikTok across time as shown in Figure 7.

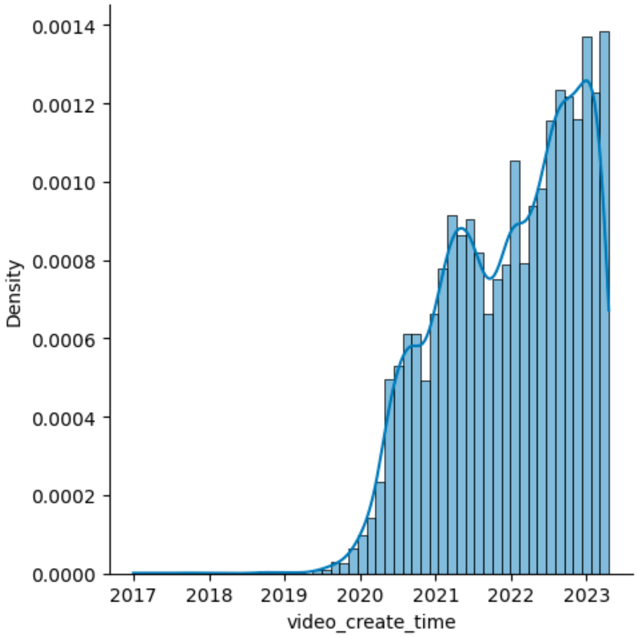


Figure 7. Number of Videos on TikTok across Time

The duration of videos is also an interesting direction for us to see. Nowadays, many hot videos on TikTok only have a few seconds, but they have tons of likes. Here's the distribution of the duration of the videos. I found that most videos are under 50 seconds, and even the peak falls around 25 seconds in duration. It follows the tendency of today's people to "short" videos, shorter and shorter.

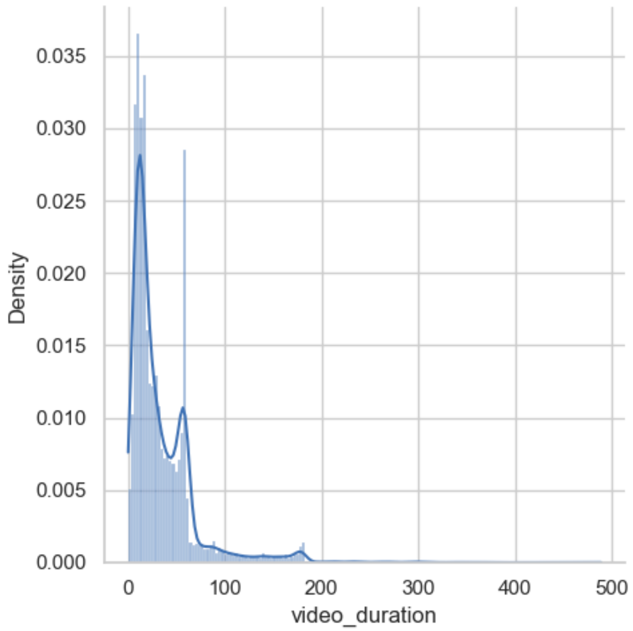


Figure 8. Duration of Videos

To take a deeper look at the videos, I'm going to split the evaluation of the videos into two parts, "quality" and "widespread".

First, I considered the number of comments, the number of likes (i.e. "diggcount" of the dataset), and the number of shares to measure the "quality" of the video. It's based on the assumption that people tend to leave a comment, hit a like button, or share it with others when they find the video itself interesting, inspiring, or educating. Even though sometimes people do share videos they don't like or complain about to others for whatever purposes, it's more uncommon than the previous scenario. So, I assumed that all people like, comment, or share it with others because of their acceptance or enjoyment of the corresponding video. And the weights of the three counting features here are 1/3 for each to simply make them equal-weighted. Hence, the equation would be:

Quality of the Video = 1/3 * (Number of Comments + Number of Likes + Number of Shares)

In this way, I filtered out 1000 top-quality videos and found the most used hashtags in them are "weightloss", "bodypositivity", "whatieatinaday" besides of hashtags like "for your page".

The result indicates that videos centered around topics such as weight loss, body positivity, and daily eating ideas tend to resonate more with audiences. This resonance is reflected in higher engagement metrics, such as likes, comments, and shares.

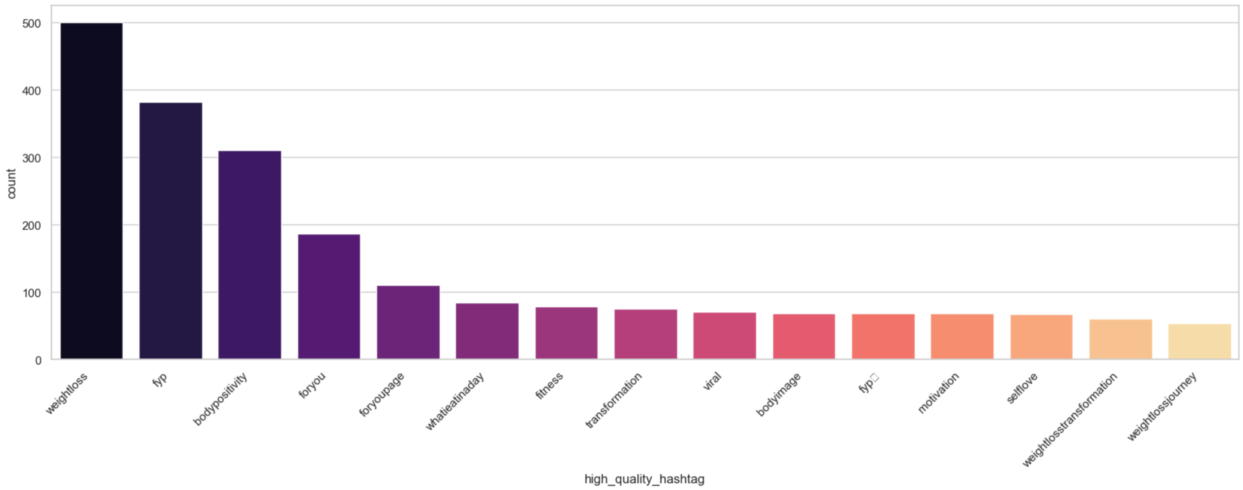


Figure 9. Most Used Hashtags in “High-Quality” Videos

Second, for the “wide-spread” metrics, I used “playcount”, the number of play times, as an alternative. With a similar process, I got the most used hashtags in top 1000 “wide-spread” videos. It mainly shows what kind of video will interest people to play, or will be recommended to more people to play. I assumed the reason people would play a video in the first place without watching any content of that video comes from a few reasons as follows: 1. They followed the video creator; 2. There are some words in the description of the video that interest them; 3. The cover of the video is attractive; 4. They are so engaged with similar videos that these videos are recommended automatically to their page. The key of the assumption is that the number of plays is not affected by the quality of the content of the video. And I got the most used hashtags shown in Figure 10.

I found that the set of hashtags is similar to the high-quality ones, but there are some new hashtags appearing, including “fitness motivation,” “loose skin,” “weight loss check,” and “and trending.” These new hashtags indicate that people like to look for motivation for weight loss through videos on TikTok. They also care about how to handle loose skin after great amounts of weight loss. Videos about “weight loss checks” or transformation are always attractive to people, whether they have or lack the motivation to lose weight. It’s reasonable that people always love watching those kinds of videos, especially when they lack motivation and do not comment or share. They just watch alone and get more motivation from those videos.

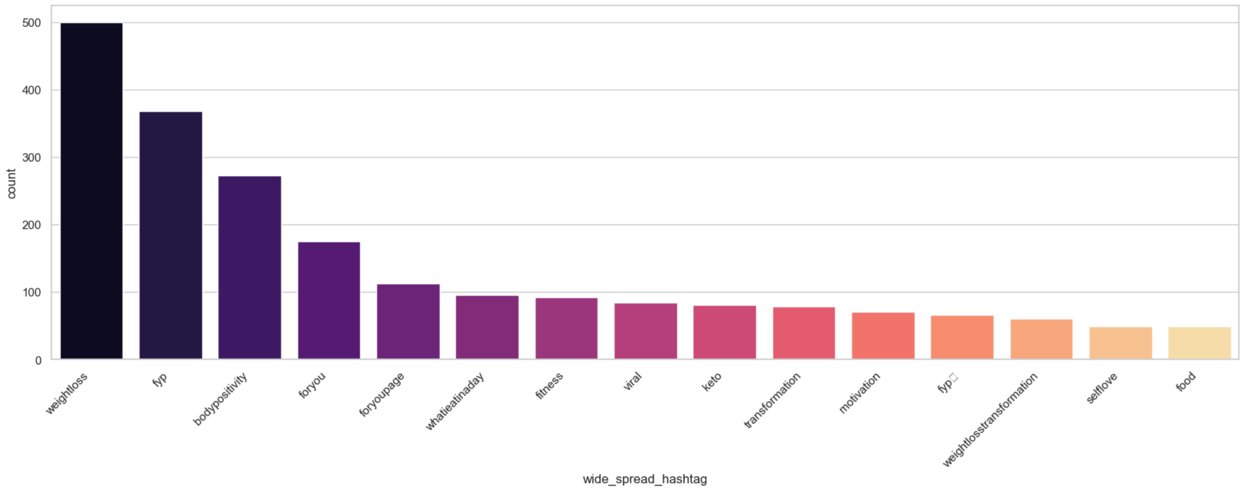


Figure 10. Most Used Hashtags in "Wide-Spread" Videos

Lastly, I combined the "quality" and "wide-spread" metrics together to get a comprehensive idea of "popular" videos. Therefore, the equation looks like:

Popularity of the Video = 1/2 * (Quality of the Video + Wide-Spread of the Video)

Similarly, I got the most used hashtags in the top 1000 "popular" videos. The results are shown in Figure 11. The "popularity" idea here is more like a blend of both "quality" and "wide-spread" as they're equally weighted.

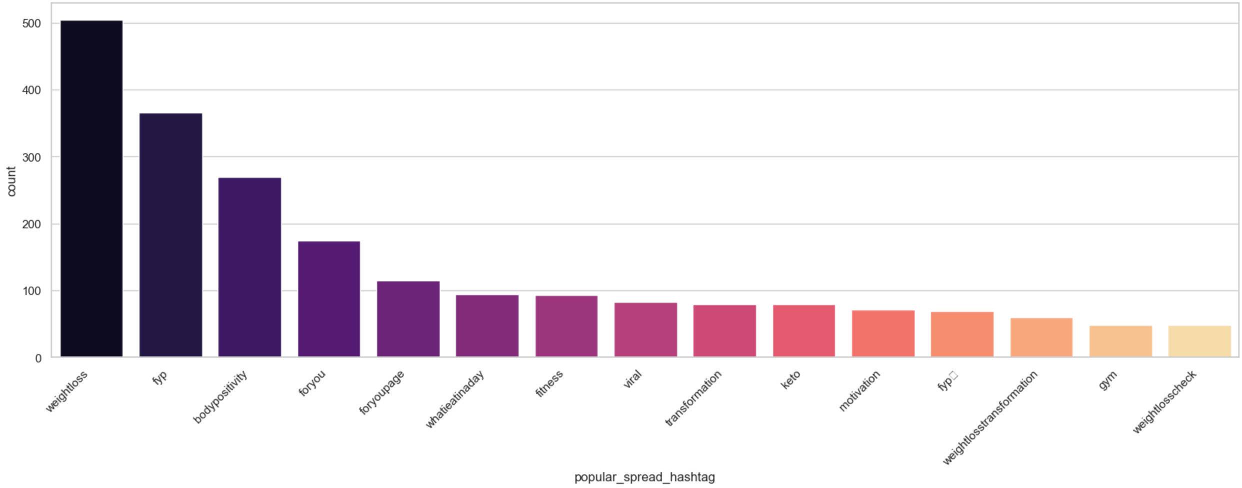
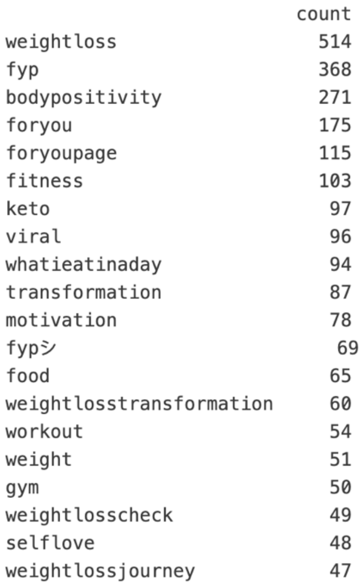


Figure 11. Most Used Hashtags in "Popular" Videos

I also found the most common "words" these popular videos like to include in their descriptions shown in Figure 12.



```
                          count
weightloss                  514
fyp                         368
bodypositivity              271
foryou                      175
foryoupage                  115
fitness                     103
keto                         97
viral                        96
whatieatinaday               94
transformation               87
motivation                   78
fypシ                         69
food                         65
weightlosstransformation     60
workout                      54
weight                       51
gym                          50
weightlosscheck              49
selflove                     48
weightlossjourney            47
```

Figure 12. Most Used Words in Popular Videos' Description

As mentioned before, the video duration is another big element that drives the popularity of the videos in a significant way. I got the top 10 common durations of these "popular" videos. From the result, I found that videos shorter than 1 minute are more likely to become popular ones. Videos shorter than 15 seconds stands out a bit more.

Table 2. Most Common Video Durations of Popular Videos

Video Duration (second)	Counts
15	96
9	51
10	50
11	45
12	44
14	42
7	42
59	35
13	32
8	31

I also took a look at the created times of those popular videos. Among the top 1000 popular videos, 367 are created in 2022, 353 are created in 2021, 172 are created in 2020, 16 are created in 2019 and none are created before 2019. There's a slight increase from 2021 to 2022.

2.2 "Comments" data

There are four main pieces of information in the comments data, including the user ID of the user who commented on the specific video, the text the user commented, the number of likes the comment received, and the user name of the user who posted the corresponding comment. The main parts I dug into were “text” and “like_count” information.

First, I filtered out the top 1000 liked comments based on the number of likes each comment received. And then, I combined the top 1000 liked comments into a single string to analyze the top words besides stop words they used, which may be part of the reason why their comments are highly likely to be liked. I got the results shown as follows.

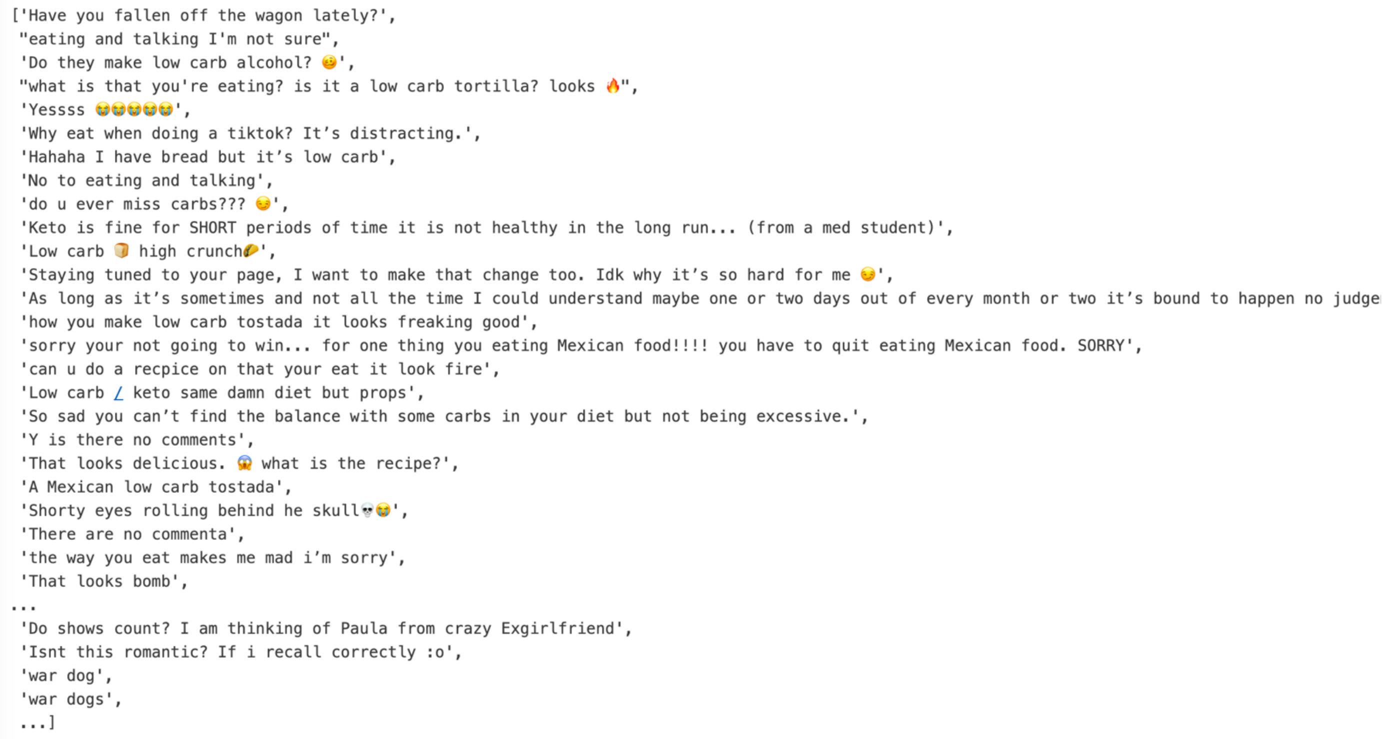
Table 3. Most Used Words in Top 1000 Liked Comments

Word	Number of Appearance
Like	43
Love	41
U	39
Proud	38
Girl	33
Thank	31
One	29
People	25
Beautiful	24
Get	24
Keep	24
Que	23
Amazing	23
Eat	22
De	22
Look	22
See	21
Feel	20
Yes	19
Got	18
Much	18
Want	18
Im	17
say	17

From the results, I found that most words in those top liked comments are positive, indicating people are more likely to like a comment with a nice vibe.

To find out the sentiment of these comments of the sample, I introduced VADER model to the dataset. At first, I tried the TextBlob Model to do the sentiment analysis. However, I tested some pieces of the comments and found the result from the Textblob Model was not that accurate, especially since it tends to give a positive estimation for some obviously negative comments. Also, one thing TextBlob was not able to handle elegantly was it could hardly analyze emojis. TextBlob is mainly designed for text analysis, so it has few capabilities in predicting the sentiment from the emojis. However, emojis are so widely used in comments on TikTok or even all social media, so it’s important to choose a model that is comfortable with playing around with emojis. Therefore, I picked up VADER instead. It’s able to understand the sentiment not only embedded in typical negations, use of contractions as negations, conventional use of word-shape to signal emphasis, sentiment-laden slang words, sentiment-laden emotions such as :) but also in UTF-8 encoded emojis.

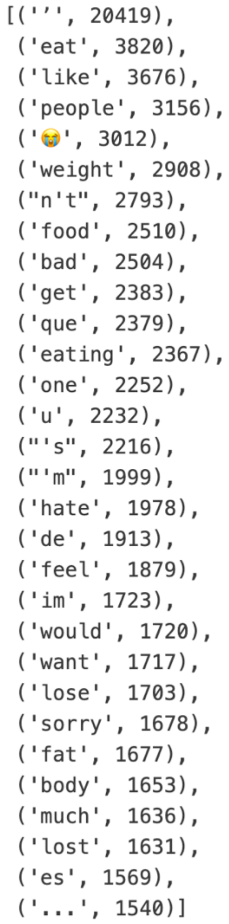
I used the VADER model to filter out the negative comments, which have compound scores less than or equal to -0.05 based on the typical thresholds. Part of the negative comments I filtered out are shown in Figure 13, and it seems the VADER model did a better job. I also calculated that the percentage of negative comments is 12.89% in this sample (there are 588,031 comments in the “comments” dataset in total, and 75,806 of them are recognized as positive.) There still exist some comments that actually were positive but were inferred as negative by the model, which could be improved more later to help enhance the performance. All the negative comments filtered out by the VADER model were saved to the “negative_comments.csv” file for further checking.



```
['Have you fallen off the wagon lately?',
 "eating and talking I'm not sure",
 'Do they make low carb alcohol? 🥴',
 "what is that you're eating? is it a low carb tortilla? looks 🔥",
 'Yessss 😭😭😭😭😭',
 'Why eat when doing a tiktok? It’s distracting.',
 'Hahaha I have bread but it’s low carb',
 'No to eating and talking',
 'do u ever miss carbs??? 😏',
 'Keto is fine for SHORT periods of time it is not healthy in the long run... (from a med student)',
 'Low carb 🍞 high crunch🌮',
 'Staying tuned to your page, I want to make that change too. Idk why it’s so hard for me 😏',
 'As long as it’s sometimes and not all the time I could understand maybe one or two days out of every month or two it’s bound to happen no judge
 'how you make low carb tostada it looks freaking good',
 'sorry your not going to win... for one thing you eating Mexican food!!!! you have to quit eating Mexican food. SORRY',
 'can u do a recpice on that your eat it look fire',
 'Low carb ∠ keto same damn diet but props',
 'So sad you can’t find the balance with some carbs in your diet but not being excessive.',
 'Y is there no comments',
 'That looks delicious. 😱 what is the recipe?',
 'A Mexican low carb tostada',
 'Shorty eyes rolling behind he skull💀😭',
 'There are no commenta',
 'the way you eat makes me mad i’m sorry',
 'That looks bomb',
...
 'Do shows count? I am thinking of Paula from crazy Exgirlfriend',
 'Isnt this romantic? If i recall correctly :o',
 'war dog',
 'war dogs',
 ...]
```

Figure 13. Some Negative Comments from All Comments

Furthermore, I explored the most used words in those negative comments. The thought comes from if we could detect these negative words accurately, then we could remind the users before they post it out as a comment. This approach is expected to enhance the overall positivity of the social media environment by filtering out negative comments. This can contribute to creating a clearer and more positive atmosphere for users engaging with the content. The top 30 used words in these negative comments are:



```
[(''', 20419),
 ('eat', 3820),
 ('like', 3676),
 ('people', 3156),
 ('😭', 3012),
 ('weight', 2908),
 ("n't", 2793),
 ('food', 2510),
 ('bad', 2504),
 ('get', 2383),
 ('que', 2379),
 ('eating', 2367),
 ('one', 2252),
 ('u', 2232),
 ("'s", 2216),
 ("'m", 1999),
 ('hate', 1978),
 ('de', 1913),
 ('feel', 1879),
 ('im', 1723),
 ('would', 1720),
 ('want', 1717),
 ('lose', 1703),
 ('sorry', 1678),
 ('fat', 1677),
 ('body', 1653),
 ('much', 1636),
 ('lost', 1631),
 ('es', 1569),
 ('...', 1540)]
```

Figure 14. Top 30 Words in Negative Comments

However, a single word can't represent everything. The whole sentence and the tone should be considered together to tell if one comment is negative.

3. Other Thoughts

3.1 Detection of Contents

3.1.1 Toxicity Detection

From the starting point of negative comments, it might be meaningful to manipulate some NLP models to detect toxicity comments or any kind of text on social media. And then remind or directly ban the content from showing, which will make the social media platform more positive and welcoming.

3.1.2 Possible Negative Contents Detection

We could detect and flag content that might be harmful to specific groups of people by analyzing data related to the video. For example, we could consider the percentage of negative comments it has, if it's over 90%, meaning that the video is likely to have some negative sides. Then, we can flag the video to remind people with certain concerns not to see it. In the long term, the ideal state would be we are able to flag one video for specific groups of people based on its content. For example, we could give videos with “pure eating” as a possible flag to people with eating disorders.

3.2 Network Science

3.2.1 Connect People Sharing the Similar Interests

People who have many common comments under the same video or videos under the same topics, it’s much more likely that these people share similar interests. For example, Alan commented on 30 videos in the past 6 months, and Joe commented on 19 videos. After looking at the videos they commented on, we might find that 5 are the same and 3 are from the same category. Then, it’s highly likely that Alan and Joe share similar interests, such as the Mediterranean diet. Therefore, they might be interested in connecting with each other, discussing and sharing videos, etc. We could also take a look at their differences. Joe may be an influencer on TikTok, while Alan is just a new user. But the bridge between them could help them connect with each other. Afterward, Alan might be interested in becoming an influencer as well. Therefore, the influence of the “Mediterranean diet” will become bigger in social media.

The reason why we care about the networks between people may partially come from the power the network could give back. Especially when caring about “positive” ideas that need to be widely spread, we need such networks to help us send our thoughts further.

3.2.2 Help People Out

To build a more harmonious, healthy social media, it’s helpful to detect contents that might have potentially negative influences on sensitive groups. For instance, it might be harmful to people who are at the edge of an eating disorder or just recovering from it to see videos about “eating too much” or “not eating at all”. However, it might be hard to directly detect these videos since they are okay with normal people who don’t have that kind of concern. So, the recommendations of these videos should be wisely built. I think it’s helpful to detect people who might have these concerns at first by analyzing their profiles, their watching habits, their commenting habits, and their own videos if there are any. After detecting those groups of people, we could pop up some warm reminders to ask them if help is needed. We can also adjust the recommendation systems to balance out the “toxic” contents to them.

The idea can also be applied to mental health. By analyzing videos users watch, comments they leave, and any texts they provide, we might be able to catch some clues that someone is in struggle. Since the internet is huge and people don’t actually know about each other, people with mental health concerns may tend to leave some thoughts online. Therefore it gives us an opportunity to identify them and then provide recommendations or available communities to them. It would probably bring some “privacy policy” things in, but as long as we just offer any possible help recommendations, it will not overstep too far.