FINESSE Modeling Challenge

Engineering & Modeling Social Data

Frank Xu
frankxu0124@gmail.com

FINESSE

Objectives & Topics

- Explore what features could be important in deciding trending posts.
  - Feature engineering
  - Feature Importance
  - Feature Interpretation

- Establish the top five trending fashion posts across Instagram in April 2020 through the 3 datasets.

- Full model architecture proposal if given more time.
  - Architecture Proposal
  - Thoughts

FINESSE

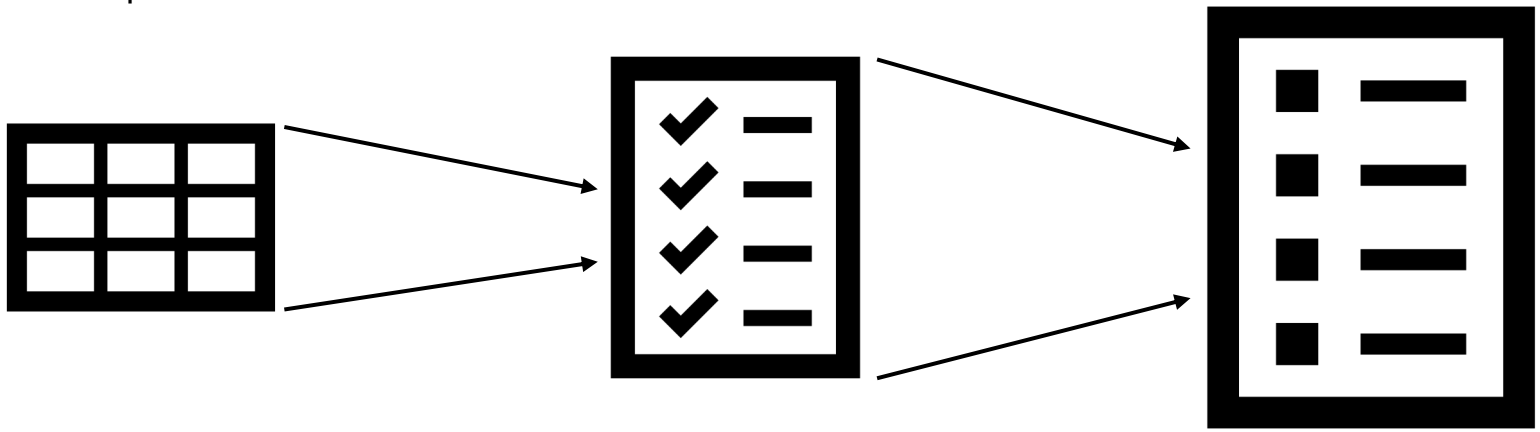
What does trending mean?

- Appearing in the "top posts" section of certain pre-filtered, trending streetwear hashtags, as described in ***hashtag_top_appearances***.
- Viewed and approved by more people, indicated by the number of likes or comments.
- There remains a lot of things unclear in this dataset, but we will be targeting in predicting **if a post could appear in top hashtag sections**.
- Logics behind reducing the problem from regression to classification:
    - Target variable is a truncated distribution.
    - With current features, regression model is not robust enough.

FINESSE

Data Overview

As mentioned below, three tables contain different number of posts. However, the posts in the first table are subset of the second one, and the second is the subset of the third. Considering the first table being the main file, we will be focusing on its 11637 posts.



post_metrics_and_comments, containing 11637 posts and main features of the posts, including comments and images.

hashtag_top_appearances, containing 25062 posts with number of days trending.

raw_post_metrics, containing 25077 posts with posts' growth over time.

FINESSE

Concerns regarding data

- All medias are saved as image (.JPG) files on AWS S3 including videos, which makes it hard to utilize the media data.
- There appeared to be some issues with comments scraping: In most of the case comments are only a subset of actual comments. However, there are situation of duplicate comments, so that the number of comments in **concatenated_comments** is bigger than **max_comments.**
- Lack of information about datasets and actual hashtags. What actual hashtags were used to collect each dataset?

FINESSE

Feature Engineering: likes & comments

First let's look at number of likes and comments. Related features appear in table 1 and 3, including max likes, max comments, and the number of likes/comments over time. **However, there are issues with those features:**

- Max likes/comments are simply the max number for each post from table 3. They should not be directly used because the days between posting and gathering data are different for each post.

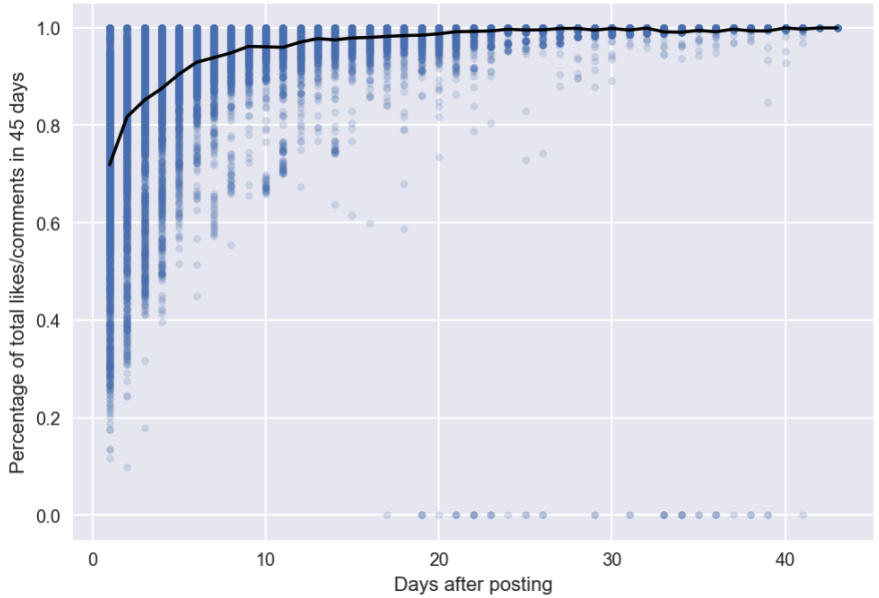
- Number of likes/comments over time could be a great feature. But still, the days between posting and gathering data are different for each post.

- **Does 4000 likes over 40 days the same as 200 likes over 2 days?**

FINESSE

Feature Engineering: likes & comments

One solution is to try to find out the distribution of how many likes/comments can a post get over time. Within the data from table 3 and a regression model, data are processed as below. In this case, we dropped posts that last less than 1 day (11637 to 10993, 5.5% difference)

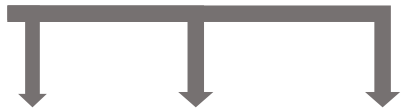


max_likes	days_after_posting	adjustment	adjusted_likes	likes_per_day (as of 45 day average)
4000	40	0.9993	4003	100
1000	10	0.9607	1041	104.1
200	2	0.8169	245	123

Example of processing likes. Same goes with comments.

Feature Engineering: actual comments

We process actual comments and only adopt last 3 new features.



Original comments	Comment counts	Emoji counts	Remove emoji, |, @, #	Translate with googletrans	Comment length (by character)	**Average comment length**	**Comment sentiment**	**Comment emoji rate**
@_o9.02 가입하자 연재야😘|아고 ㅠㅠㅠㅠ 힘드실텐데 아자아자 화이팅입니다 !!🙏|@addstyletome_ 화이팅입니다:)|@59seok 😭:(힘내시죠|선생님 상의정보좀 알수있을까요😃	5	4	가입하자 연재야 아고 ㅠㅠㅠㅠ 힘드실텐데 아자아자 화이팅입니다 !! 화이팅입니다:) :(힘내시죠 선생님 상의정보좀 알수있을까요	When you sign up, you'll know it's Yeonjae. This is fighting	60	**12.8**	**-0.3612**	**0.0625**
She is cute 💖💖|So beautiful|#GirlDad|💯🔥|Awww she is beautiful! Young Queen	5	5	She is cute So beautiful Awww she is beautiful! Young Queen	She is cute So beautiful Awww she is beautiful! Young Queen	59	**12.8**	**0.9132**	**0.0781**

FINESSE

Feature Engineering: other features

Comment-like ratio	Commenting generally takes more time than liking. A higher comment-like ratio might suggest higher user interactive rate.
Weekday	People tend to have different behaviors on different part of a week.
Image count	Count how many images/videos are included in one post.
Post frequency	Calculate the frequency of posts for each user that appears in the dataset
Trending	An Indicator of whether the post has been on on top trending page.
Following	Following count for each user. From original data.
Followers	Followers count for each user. From original data.

FINESSE

Feature Engineering: username

Username is a string with great predicting power in this case. We are going to **target encode** username, so that our model can utilize this information.

username	trending
lukebhuby	0
d.ave.y	1
lukebhuby	1
d.ave.y	0
d.ave.y	1

Target Encoding: Taking the average of the target variable for each class.

username	trending
0.50	0
0.67	1
0.50	1
0.67	0
0.67	1

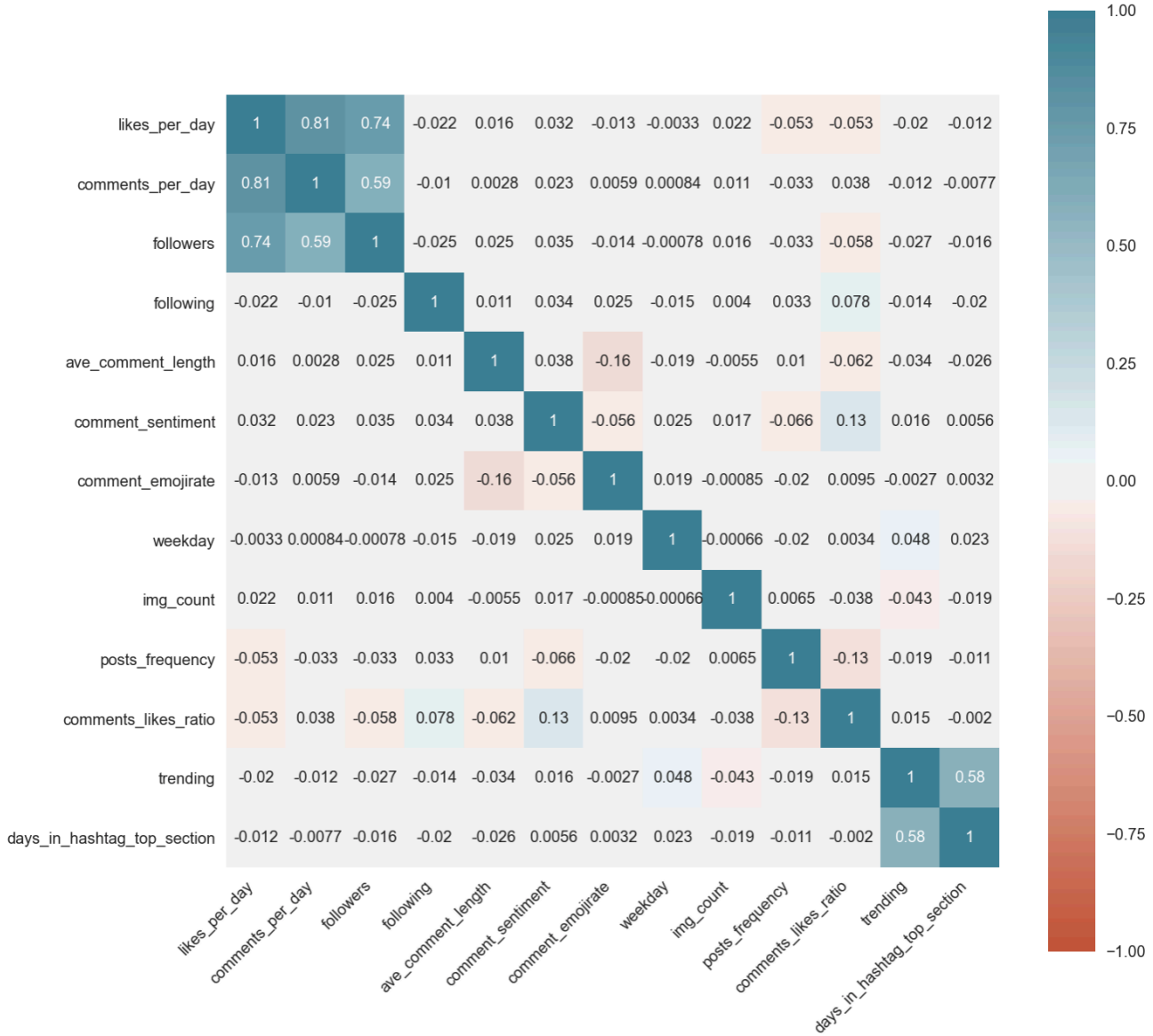
FINESSE

Feature Engineering: features abandoned

- Max likes and comments. These are replaced by likes and comments per day, as mentioned before.
- Post URL. Serves no predicting power.
- Max views. Only 5.2% of posts include videos. Can be partially replaced by likes & comments.
- Caption. Some of the captions are shortened.
- Images. Due to limit time and lack of computation power at hand.

Feature Engineering

- Likes, comments per day and followers are highly correlated.
- Trending is reduced from days in hashtag top section, thus the high correlation.
- Other features show no strong correlation with trending or days in hashtag top section.



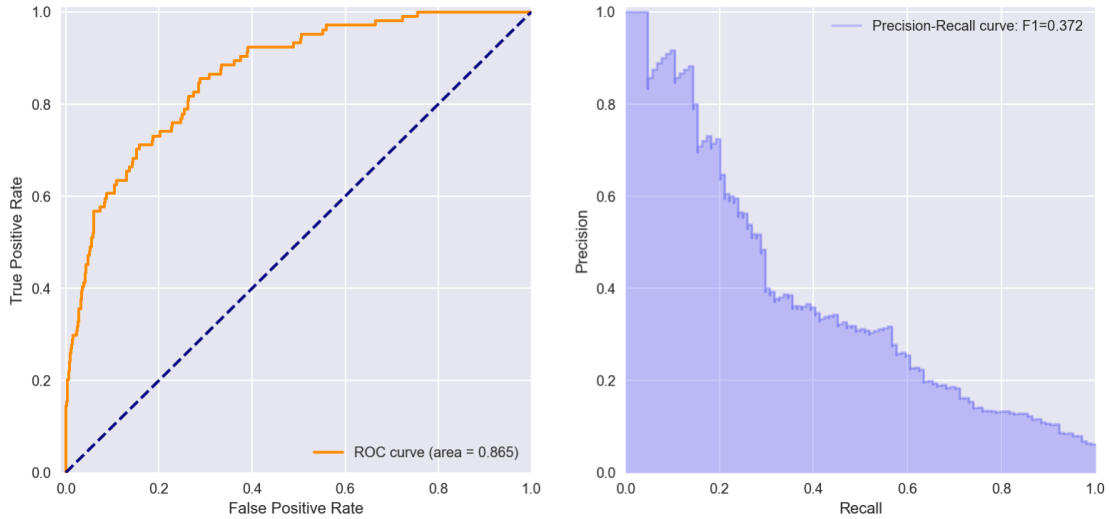
FINESSE

Modeling

username
likes_per_day
comments_per_day
comment_like_ratio
following
followers
average_comment_length
comment_sentiment
comment_emoji_rate
weekday
post_frequency
img_count

Tuned
LightGBM
Classifier

trending

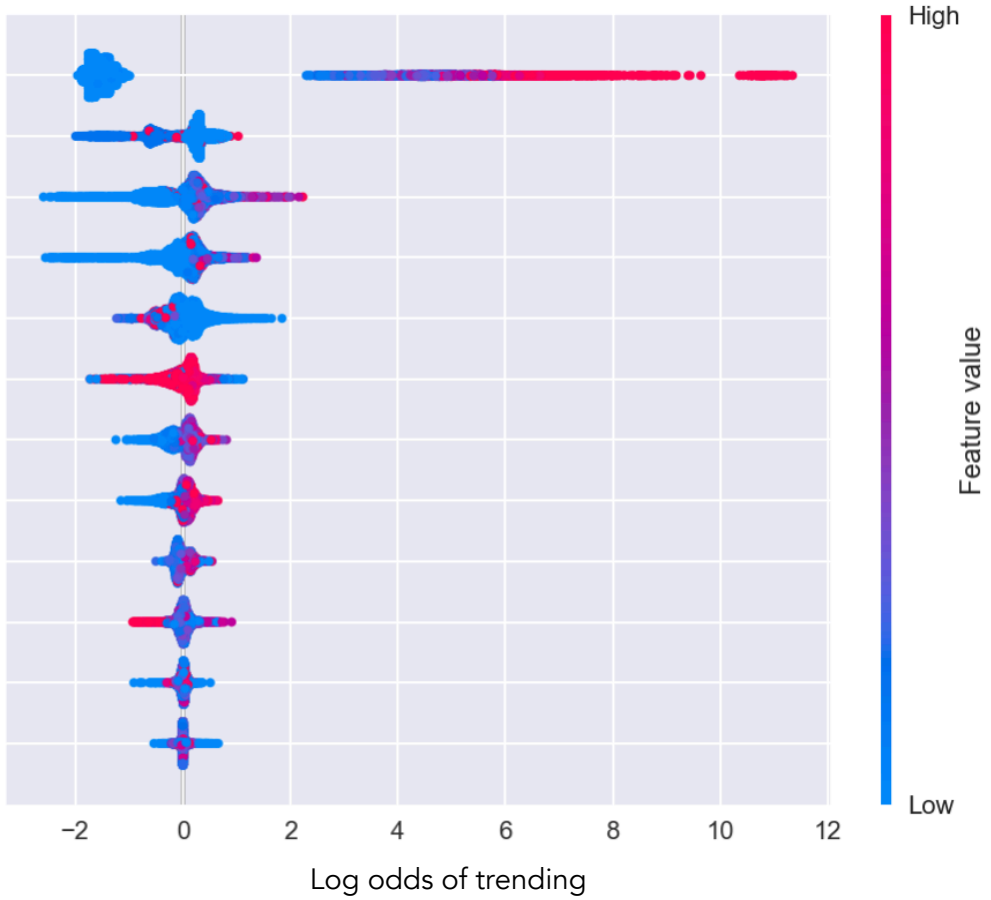


Model Performance

Model Interpretation

- We got the feature importance and feature influence with Shap values of our model.
- Higher importance does not indicate stronger causal relations.

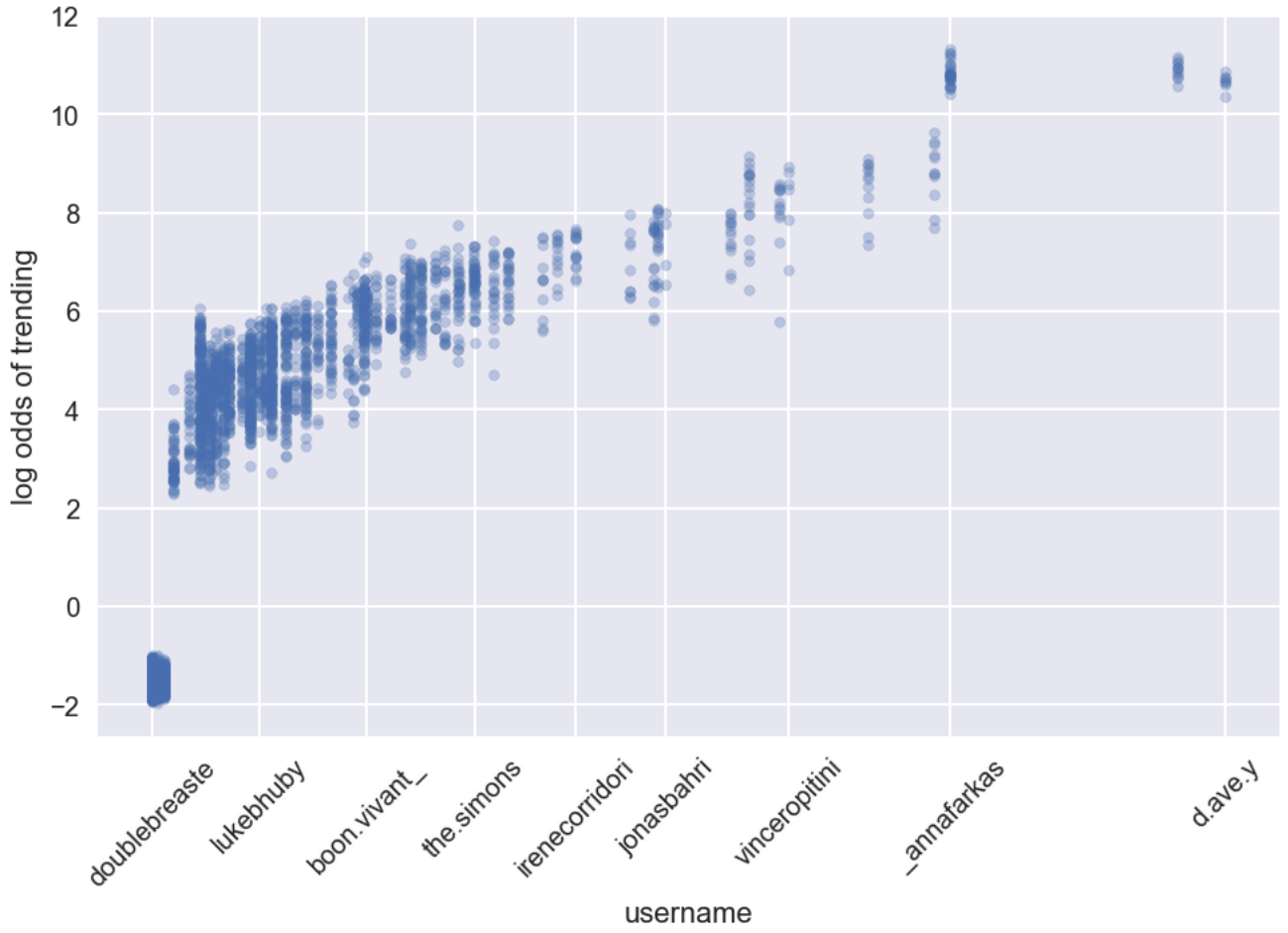
Feature	Importance
username	2.392521
img_count	0.421534
comments_per_day	0.403657
likes_per_day	0.267306
followers	0.242641
comment_sentiment	0.204935
ave_comment_length	0.183090
weekday	0.142012
following	0.107972
comment_emojirate	0.097686
comments_likes_ratio	0.051037
posts_frequency	0.038459



Model Interpretation

- Username has most predicting power in this case. Here is a dependent plot of username and log odds of trending.

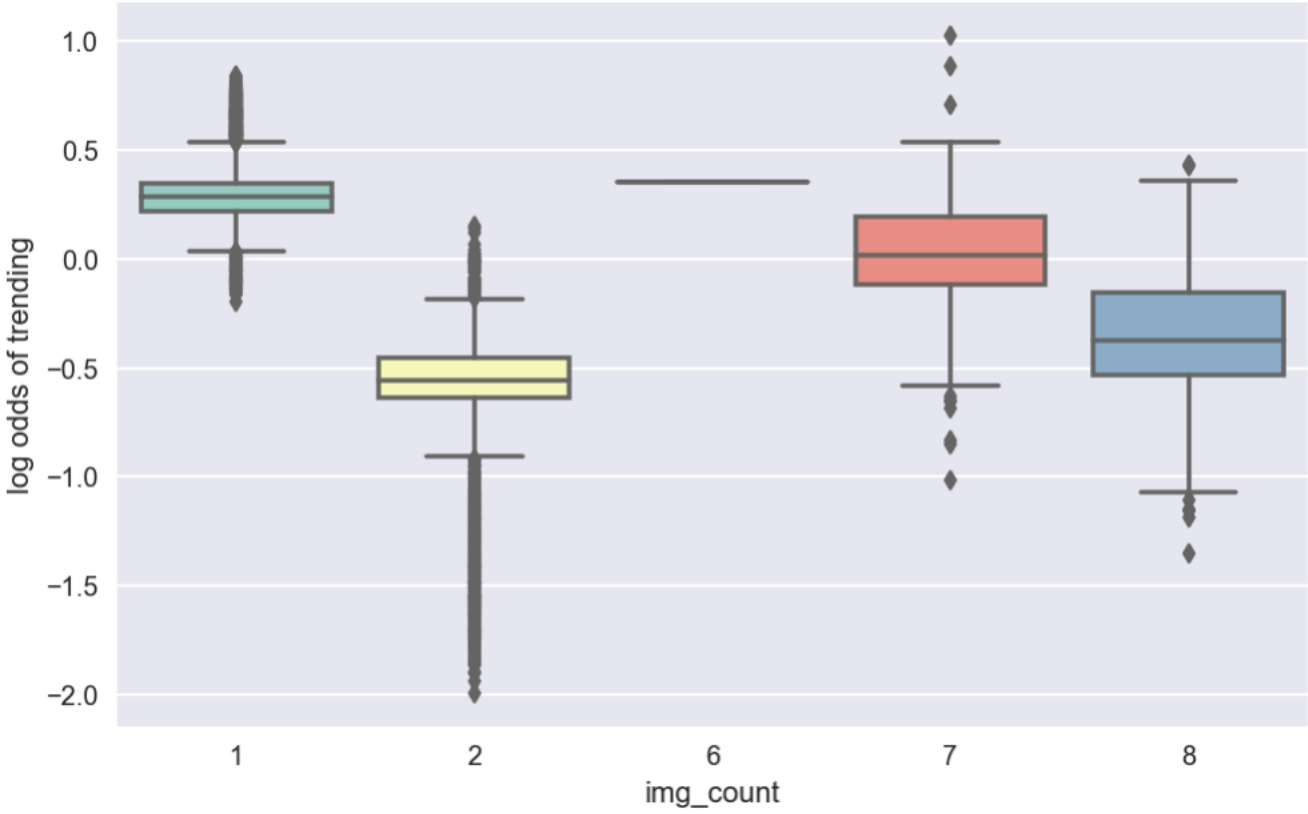
- From this dataset, we can see posts from **@d.ave.y** are almost 60000 (exp(11) = 59874) times more likely to be trending than the average level. On the opposite, posts from **@doublebreaste** are about 1/7 (exp(-2) = 0.14) of the average level.



FINESSE

Model Interpretation

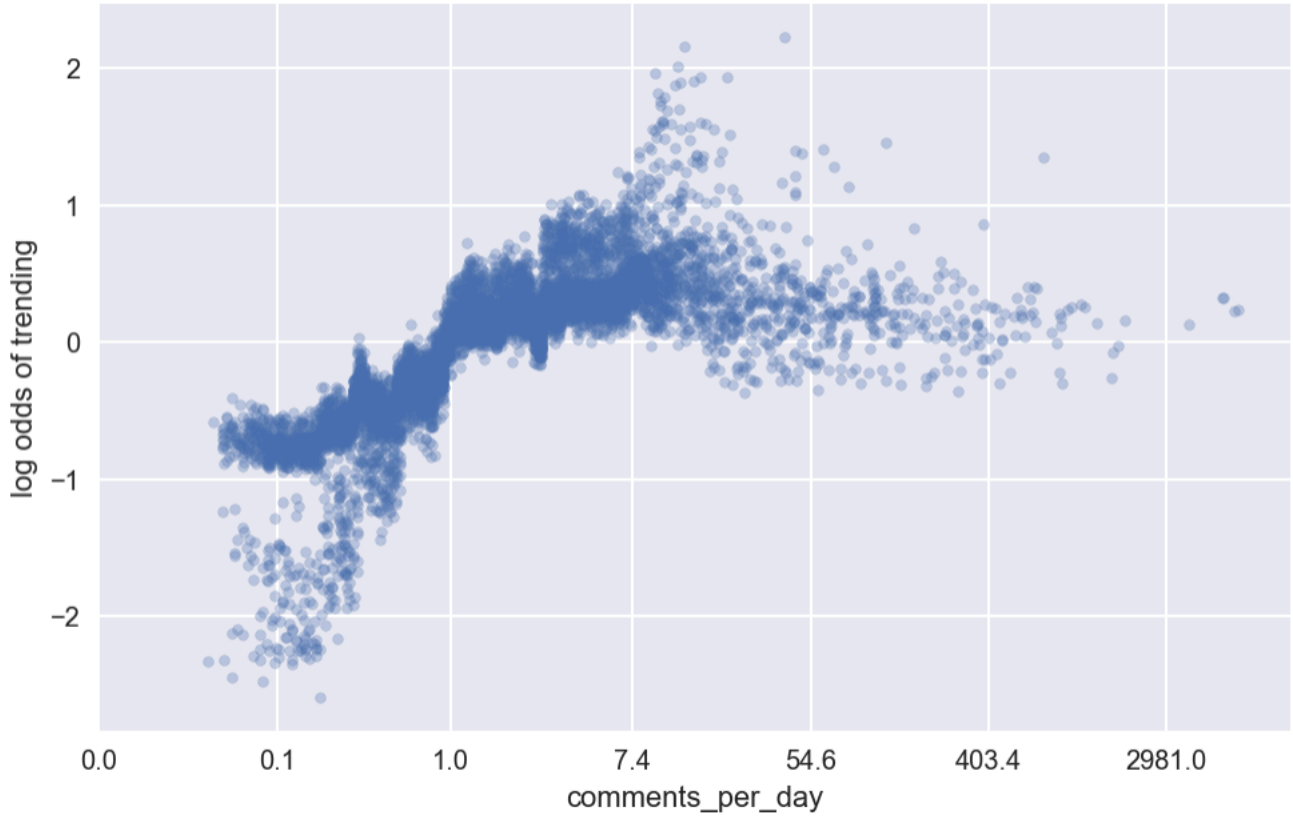
- Image count has 2nd most predicting power in this case.
- From this dataset, we can see posts with only 1 image/video are 35% (exp(0.3) = 1.35) more likely to be trending than the average level. On the opposite, posts with 2 images/videos are 45% (exp(-0.6) = 0.55) below the average level.



FINESSE

Model Interpretation

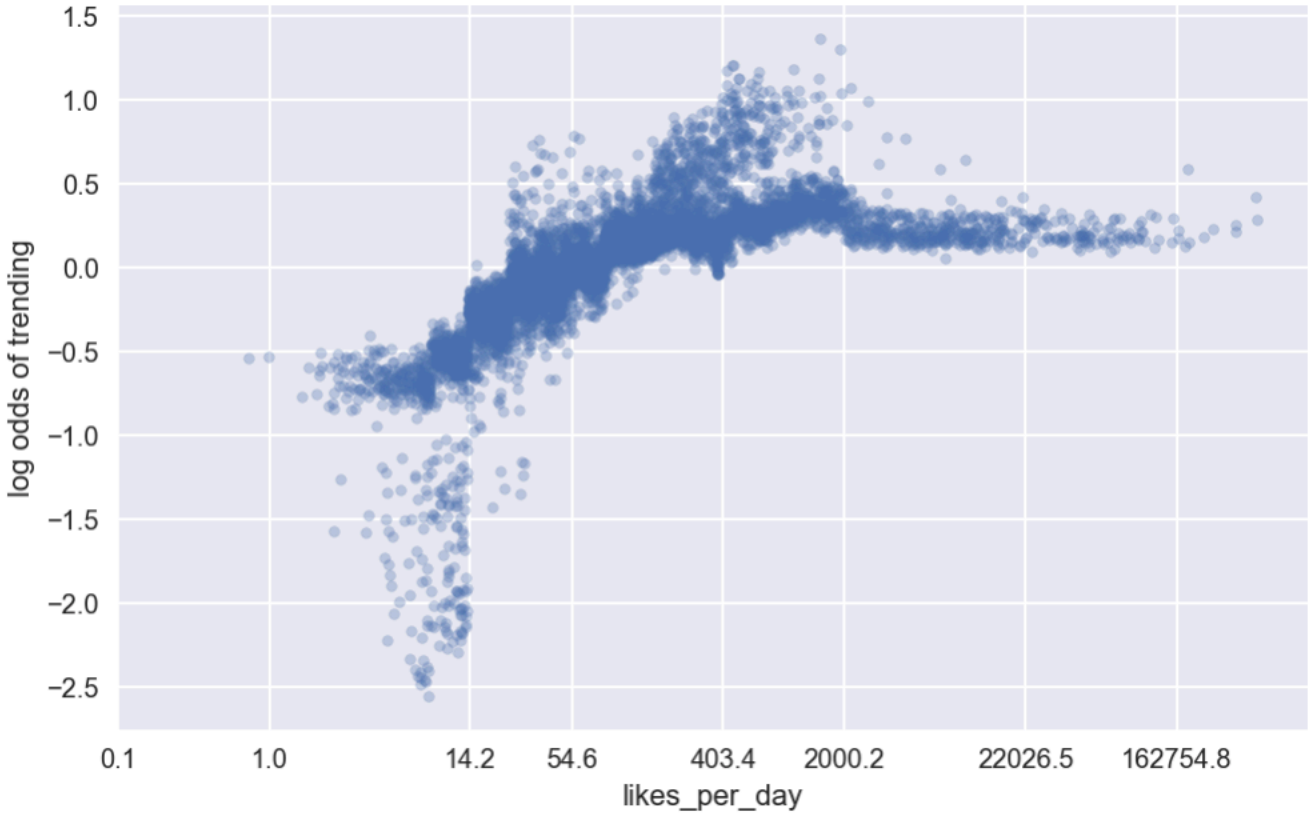
- Comments per day has 3rd most predicting power in this case.
- From this dataset, we can see posts with more than 1 comment per day are more likely to trend, and those less than 1 are less than average. Posts with 5~30 comments per day are most likely to trend.



FINESSE

Model Interpretation

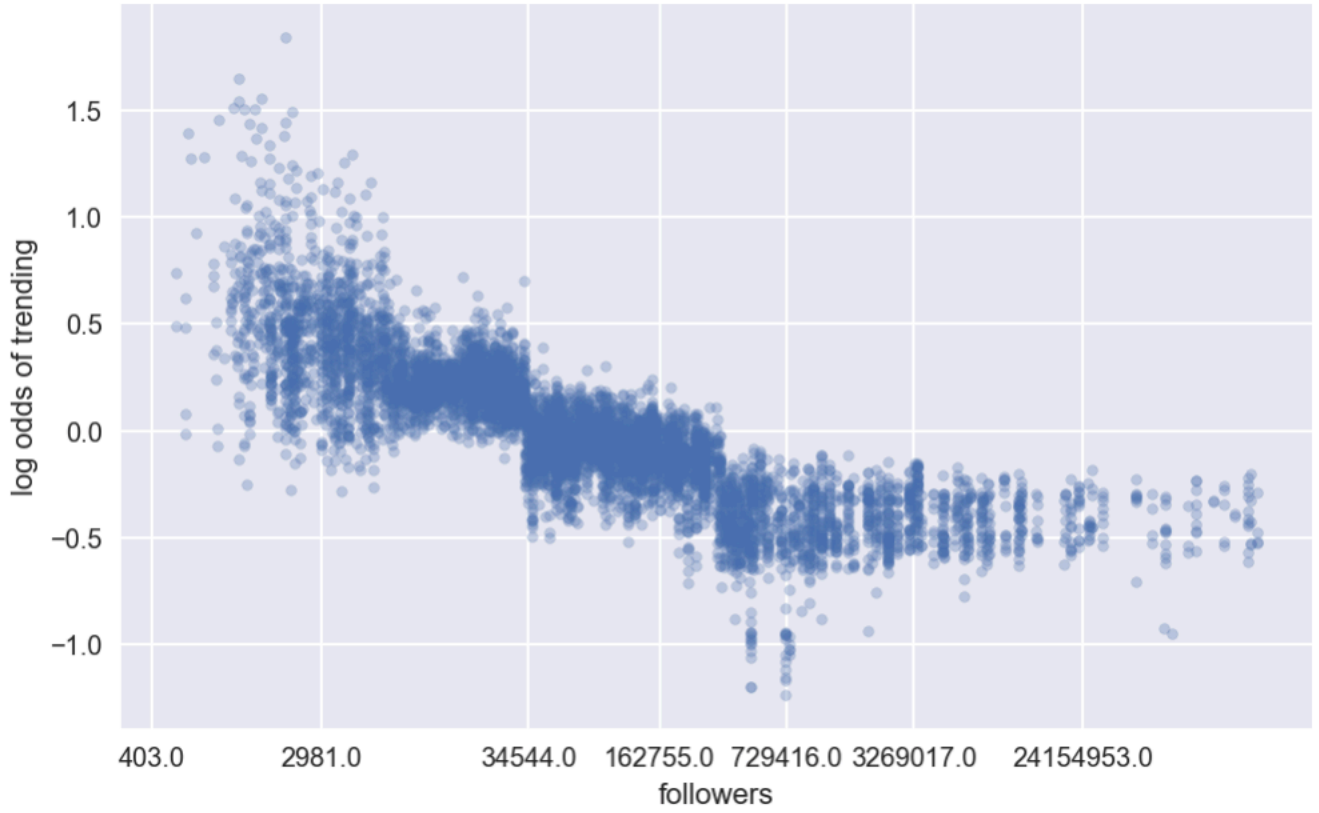
- Likes per day has 4th most predicting power in this case.
- From this dataset, we can see posts with more than 55 likes per day are more likely to trend, and those less than 14 are at least 40% less than average. Posts with 100~2000 likes per day are most likely to trend.



FINESSE

Model Interpretation

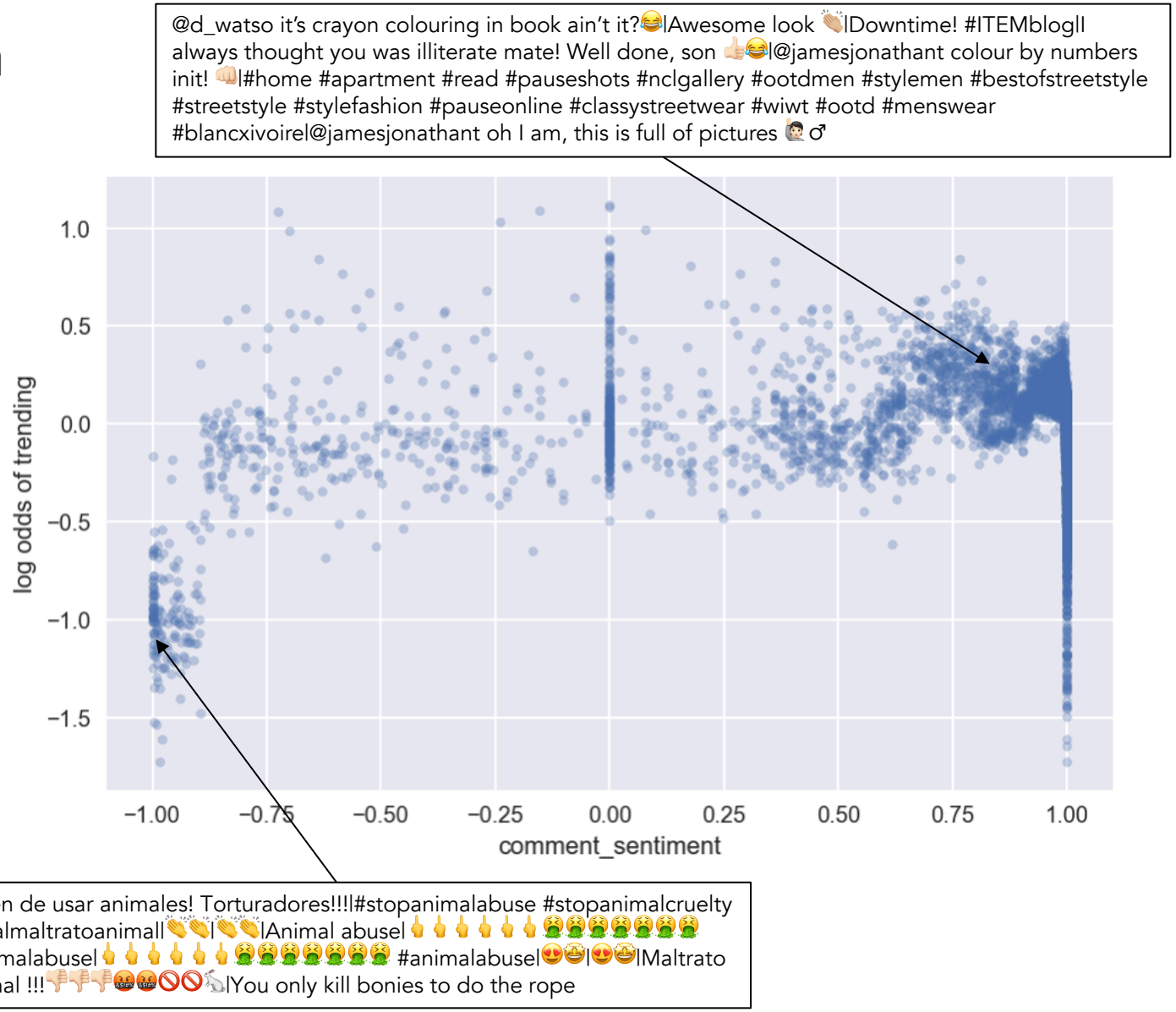
- Followers has 5th most predicting power in this case.
- Our algorithm seems to favor users with less than 35k followers. And if you have more than 1M followers, your posts are 40% less likely to be trending.



FINESSE

Model Interpretation

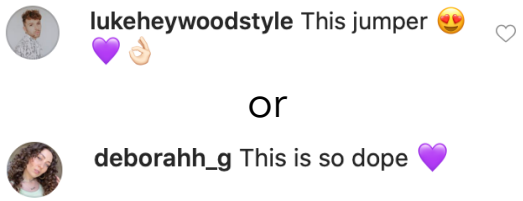
- Comment sentiment has 6th most predicting power in this case.
- Posts with comment sentiment score higher than 0.6 are more likely to be trending than average level.
- If your post spreads horror or discomfort, rest assure it won't be trending.



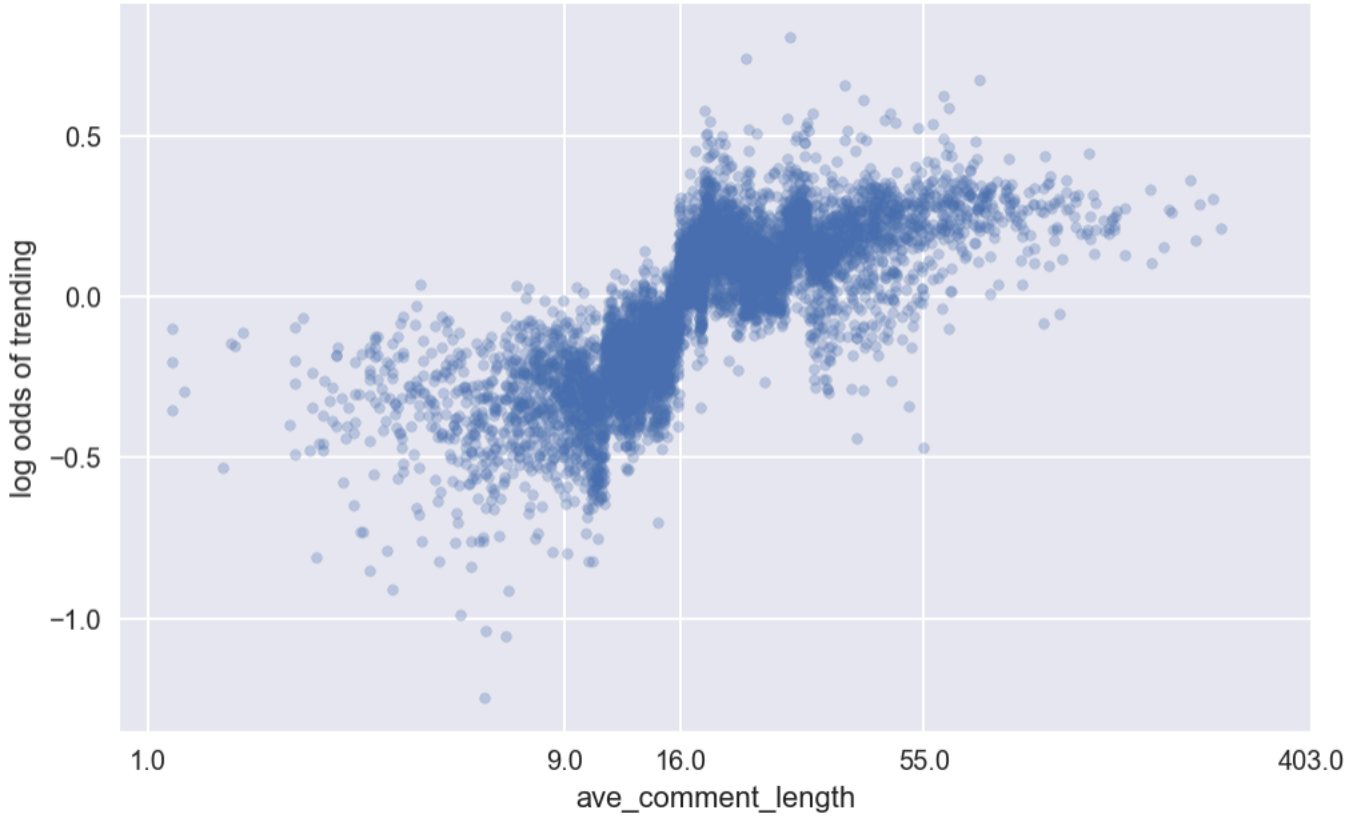
FINESSE

Model Interpretation

- Average comment length has 7th most predicting power in this case.
- Posts with average comment length bigger than 16 characters (including spaces and emojis) are more likely to be trending than average.
- What does a 16-character comment look like?



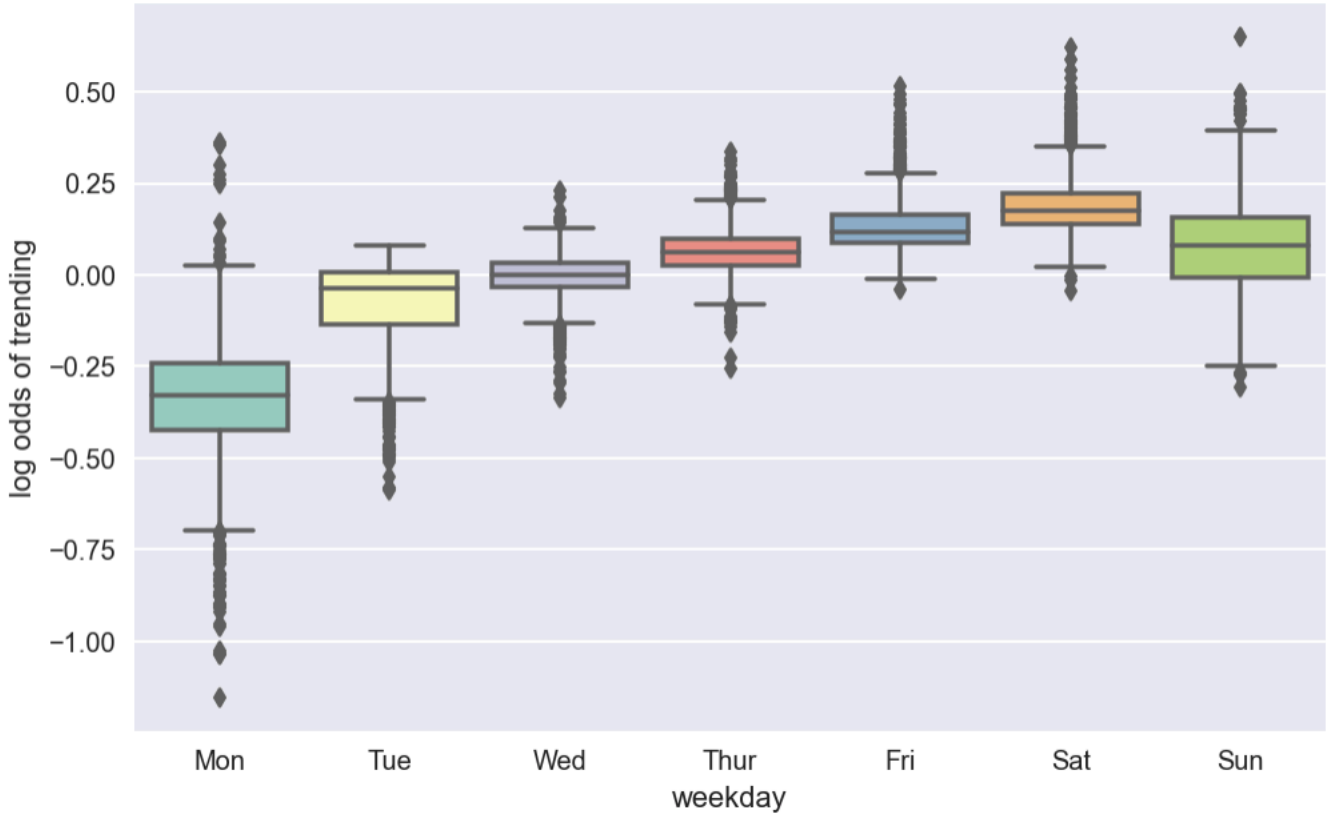
or



FINESSE

Model Interpretation

- Weekday has 8th most predicting power in this case.
- Posts posted on the latter half of a week are more likely to be trending.
- Why Monday? Because it's the day that comes after Sunday and ends your glorious weekend.



FINESSE

Top 5 trending posts

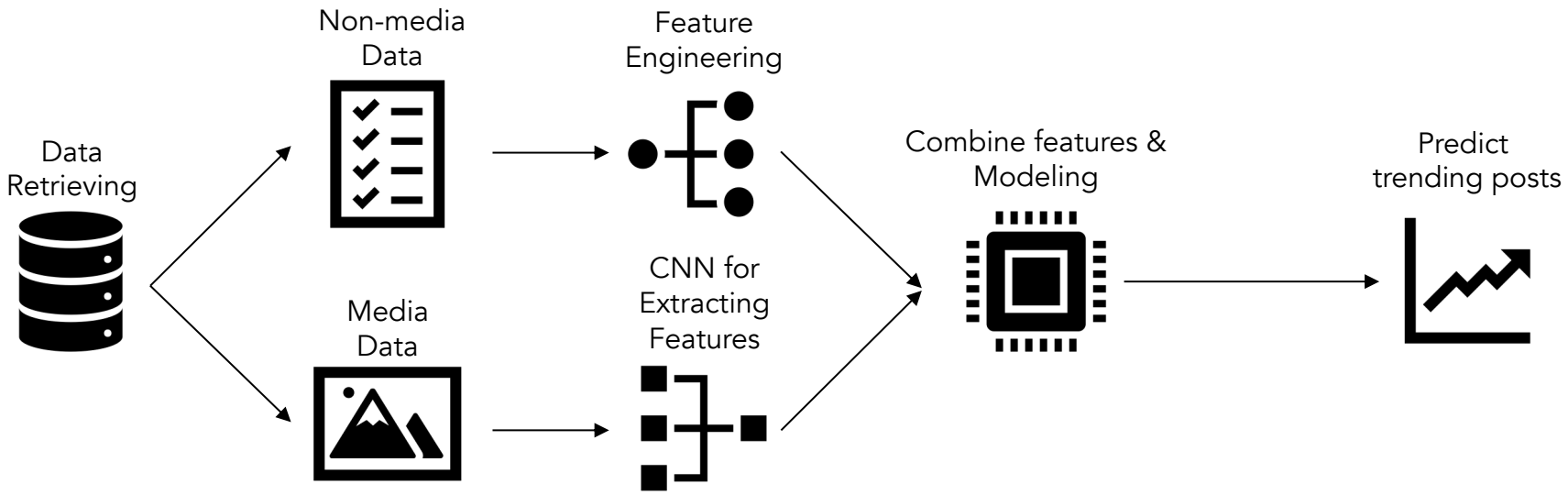
Top 5 posts that are most likely to be trending according to the model trained are:

post URLs	username	followers	days in hashtag top section
https://www.instagram.com/p/B_XUYo4KZN3/	outfitsofmunich	1037	5
https://www.instagram.com/p/B_X1YtqKVOD/	biggerswoosh	1295	2
https://www.instagram.com/p/B_XdCntH7QU/	kauzenshop	61200	4
https://www.instagram.com/p/B_ahhGTKWPW/	outfitsofmunich	1037	6
https://www.instagram.com/p/B_ajoxYo_t8/	jason.kural	1156	2

For more information please check out the notebook file

P.S. I also gathered the top 5 post that are labeled as not trending right now, but most likely to be trending according to the model, which is included in the notebook file.

Full Model Architecture



FINESSE

Full Model Architecture

Future directions:

- More NLP features:
  - **Key words:** potential topics might influence trending on Instagram.
- Emojis:
  - **Sentiment score of emojis**: people use more and more emojis everyday.
- Images:
  - **Extract features directly using a CNN from images**: features like textures and shapes can be extracted and combined with other features for the main model.
  - **Using a pre-trained model to identify main subject of the image**: different post subjects have different feedbacks.
- Extra data:
  - **Hours of day posting**: works just like weekdays.
  - **Account types**: Is the account business or not? Is it labeled as celebrity or creator?
  - **Profile website link types**: whether the user has a link, and whether the link goes to YouTube or a personal blog.
  - **Profile pictures**: can be processed together with post images for each post.
  - **Instagram stories data**: statistics show that stories account for 1/3 of Instagram business & activities for now. This might be a whole new project itself.

Thoughts

- First of all, thank you for bringing this amazing project. It's been one of the most fun and productive weeks that I have ever had.

- There exists a lot of projects predicting Instagram likes, and it's a lot easier. Trending, on the other hand, is mixed more with human reactions and a more complicated process, which granted it more randomness and unpredictability.

- The goal of this project is worth thinking about. My model shows that Instagram tend to favor small to medium sized accounts with top section in hashtags. However, larger accounts do influence more people.