# [Data] Characterizing Clickbaits on Instagram

Welcome to our data project page. Our Instagram dataset is composed of information about 7,769 posts on Instagram. The data collection was done over a two-week period in July 2017 using an InstaLooter API. We searched for posts mentioning 62 internationally renowned fashion brand names as hashtag, which are listed below.

| Brand name |
| --- |
| abercrombie, acnestudios. alaia, alexandermcqueen, alexanderwang, americanapparel, americaneagle, balenciaga, balmain, brioni, brunellocucinelli, burberry, calvinklein, cartier, celine, cesareattolini, chanel, chloe, cos, dolceandgabbana, dsquared2, dvf, ermenegildozegna, fendi, forever21, gap, givenchy, goyard, gucci, hermes, hm, hollister, isabelmarant, jcrew, jilsander, katespade, kenzo, kiton, lanvin, loropiana, louisvuitton, maisonmargiela, marcbymarcjacobs, marcjacobs, marni, michaelkors, miumiu, moncler, nancygonzalez, paulsmith, rebeccaminkoff, saintlaurent, sandro, stellamccartney,thombrowne, tomford, topshop, toryburch, uniqlo, valentino, vetements, viviennewestwood, zara |

Our data contains a label indicating whether a given post is a clickbait or not. Also, we contain information associated with text as well as image features based on deep learning (Convolutional Neural Network, CNN). The list of learned features includes generic visual features (representing common visual traits) and semantic image features (domain of fashion). Please refer to our paper for the full description of how we built our deep learning model.

## Papers

- Yu-I Ha, Jeongmin Kim, Donghyun Won, Meeyoung Cha, Jungseock Joo. Characterizing Clickbaits on Instagram. In Proc. of the International AAAI Conference on Web and Social Media (ICWSM), June 2018

## Data Characteristics

Below describe the data fields, followed by an example. Please note that, in concern for privacy, we anonymized any mentions of user screen names that appear as @*** (instead of @username). We also do not share the full image content but give the image URL in case they are public posts.

## *1. Basic information about posting user and each post. (Meta features)*

- ***Link:*** A web URL (if any) contained in each post
- ***Image URL:*** A web link to the image file of the post
- ***Likes:*** The total number of likes per post
- ***Comments:*** The total number of comments per post
- ***Followings:*** The number other users a given uploader is following (i.e., distinct sources of the uploader)
- ***Followers:*** The number of other users subscribing to the uploader's account (i.e., fans of the uploader)
- ***Media Counts:*** The number of total posts contributed by the uploader
- ***Searched Tag:*** A brand name used in post-search process, used as a hashtag in user's post
- ***Location Existence:*** Whether contains a location tag in each post
- ***Clickbait:*** Whether the post labeled as clickbait or not (*0=clickbait, 1=non-clickbait)*
  * We evaluated Instagram posts to label the discrepancies between images and the accompanying text (searched hashtag, brand name) by employing annotators at the Crowd Flower (www.crowdflower.com).

## *2. The textual content variables and descriptions. (Text features)*

- ***Hashtags:*** The list of hashtags in the caption
- ***Captions:*** Text description of a post provided by the user, including the hashtag
- ***Length of hashtags:*** The total string length of the hashtag
- ***Length of captions:*** The total string length of the caption
- ***URL inclusion:*** Whether contains URL (i.e., www.~~.com) in the caption
- ***Mention inclusion:*** Whether contains mentions (@username) in the caption
- ***Emoji count:*** The total number of emoji per post
- ***Emoji Existence:*** Whether contains Emoji in the caption
- ***Emoji Portion:*** The total proportion of emoji to the length of each posting's caption
- ***Top 100 hash of Insta:*** The usage count of top 100 popular hashtags within Instagram
- ***Top 100 hash within data:*** The usage count of top 100 mentioned hashtag within our data
- ***Top 100 co-mentioned hash pair:*** The usage count of top 100 co-mentioned hashtag pair within our data

## *3. The visual content variables and descriptions (Image features)*

** Selfie~Outdoor: probability*

- ***Selfie:*** Whether contains a face that occupies 50% of the height
- ***Body_snap:*** Whether contains any body part
- ***Marketing:*** Whether contains runway or ceremony scenes
- ***Product-only***. Whether contains products without persons
- ***Non-fashion:*** Whether missing fashion products
- ***Face:*** Whether contains frontal or side faces

- ***Logo:*** Whether contains any brand logo
- ***Brand Logo:*** Whether contains specific brand logo
- ***Smile:*** Whether contains any smiling faces
- ***Outdoor:*** Whether contains outdoor background
- ***0~2047 (2048-dimension):*** The output of the last feature layer (2048-dimensions) generated from an *ImageNet* pre-trained model (Krizhevsky, Sutskever, and Hinton 2012).

**[Data Example]**

Credit by @ mrs_elisi

Field name:
Link, ImageURL, Likes, Comments, Followings, Followers, MediaCounts, SearchedTag, LocationExistence, Clickbait, Hashtags, Captions, LengthOfHashtags, LengthOfCaptions, URLInclusion, MentionInclusion, EmojiCount, EmojiExistence, EmojiPortion, Top100HashOfInsta, Top100HashWithinData, Top100ComentionedHashPair, Selfie, BodySnap, Marketing, ProductOnly, NonFashion, Face, Logo, BrandLogo, Smile, Outdoor, 0~2047

Data:
https://www.instagram.com/p/BSrEjhFDAqe/, https://instagram.ficn2-1.fna.fbcdn.net/t51.2885-15/e35/17819286_1357748670971275_1112622878735466496_n.jpg, 8, 0, 517, 70, 91, jilsander, 0, 1, #loveofmylife #parfum #jilsander #sun #forwomen #formen #loveit #soriechtdersommer #neuerlieblingsduft #hmmm, Ab heute mit Partnerduft unterwegs Danke Schatz #loveofmylife #parfum #jilsander #sun #forwomen #formen #loveit #soriechtdersommer #neuerlieblingsduft #hmmm, 108, 156, 0, 0, 2, 1, 0.012658228, 1, 3, 2, 0.00206, 0.00637, 0.0119, 0.00502, 0.989, 0.00784,0.999, 0.102, 0.00458, 0.00124, 1.129934311, 1.786201, 1.461385131,... (The rest is omitted)