Data analysis | Kira Schacht | 22.05.2020

Undress or fail: Does Instagram favour posts that show more skin?

Table of Contents

Data collection 2

Data analysis 2

Indicators: Raciness and nudity 2

Raciness 2

Nudity 3

Metric: Share in encounters vs shares in posts 3

Statistical hypothesis testing 4

Alternatives explored 4

Share in encountered posts vs shares in created posts 4

Share of possible audience 4

Results 5

Raciness 5

Nudity 5

Individual donors 6

Limitations 6

Summary 7

Instagram is the EU's second-largest social network with around 140 million monthly users. It is used by many professionals to attract audiences and reach customers. The success of their content has a direct impact on their professional lives.

The platform ranks posts in users' newsfeeds according to what "users care about most". This may be influenced, for instance, by what has attracted attention in the past. If many users like pictures of scantily clad women, the Instagram algorithm may present pictures of this kind more prominently, leading to more interactions and perpetuating the effect. This would create an incentive for content creators to show more skin in their posts in order to reach a larger audience.

In this data analysis, we explore the hypothesis that pictures of women perform better if they have less clothing on, and that the same is not true for men. You can find the code behind this analysis on GitHub.

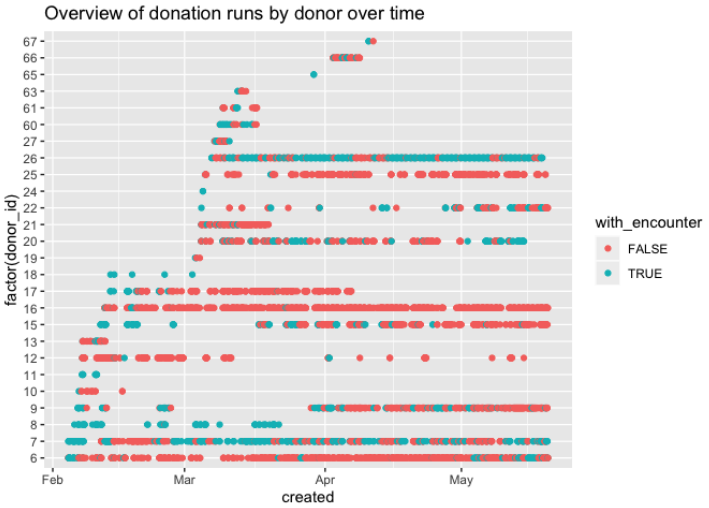
Data collection

We selected a sample of 46 IG users who derive income from their IG presence, such as providers of services, operators of an online shop or users who sell sponsored posts. They are referred to as “monitored accounts” in our context. The sample contains 31 accounts run by women and 15 accounts run by men. We analysed the content of each photo posted by monitored accounts using the Google Vision API.

A sample of 31 regular users (data donors) installed a Firefox add-on and followed 3 monitored users each. The add-on scrapes the newsfeed of data donors at regular intervals. During each donation run, it views the selection of posts presented at the top of the newsfeed (usually 12 posts at a time). When an item from a monitored user appears in these items, information about the encounter is sent to a database. All information about data donors is anonymized.

Between 31 January and 20 May 2020, we collected information on 2825 posts by monitored accounts containing a total of 3351 images. For the analysis, we considered all posts with images by the 41 monitored users who were followed by at least one data donor. This amounted to a total of 1831 posts containing 2886 images.

Within our analysis time frame, the 31 data donors made a total of 6320 donations, logging 3177 encounters with 549 individual posts by monitored accounts. 5 donors didn’t make any data donations. The scale and frequency of donations differed by donor.



Data analysis

We analysed the data collected from scans of monitored accounts and newsfeed donation runs.

Indicators: Raciness and nudity

To find out whether posts contained images of people with few clothes on, we used two indicators generated from the Google Vision API image analysis: “raciness” and “nudity”.

Raciness

For each picture, the Vision API returns a safe search rating indicating whether or not a picture contained “racy” content. The feature is measured on an ordinal scale with the

possible values `VERY_LIKELY`, `LIKELY`, `POSSIBLE`, `UNLIKELY` and `VERY_UNLIKELY`.

Racy, in this context, refers to sexually suggestive content like "skimpy or sheer clothing, strategically covered nudity, lewd or provocative poses, or close-ups of sensitive body areas."

For the purposes of our analysis, a picture labelled racy is one that received a raciness rating of either `VERY_LIKELY` or `LIKELY`. A non-racy picture is one rated either `UNLIKELY` or `VERY_UNLIKELY` racy. Images marked `POSSIBLE` are labelled as undecided.

Nudity

The Vision API also returns a collection of labels that describe the content of each picture (e.g. `Landscape, Vacation, Window, Thigh`).

To complement the safe search rating, we manually compiled a list of labels indicating nudity. We started by analysing which labels are most often associated with raciness to inform this list. These were mostly labels describing body parts, underwear or swimwear.

For all labels tagged in more than 50 images, we then manually noted whether they indicated nudity, using samples of images to test our judgement. The relevant labels were:

> `Abdomen, Barechested, Bikini, Bodybuilding, Brassiere, Chest, Lingerie, Muscle, Skin, Stomach, Swimwear, Thigh, Trunk, Undergarment, Waist.`

Since we were interested in exploring possible gender differences in the way Instagram's algorithms treat nudity, we filtered these labels for ones that are associated with one user gender in at least 90 % of pictures. We adjusted the resulting list so that only clearly gendered terms remained. The final list of labels indicating gendered nudity was:

> Women: `Brassiere, Lingerie, Undergarment, Bikini`
> Men: `Barechested, Bodybuilder`

Metric: Comparison of encounters vs posts

If Instagram's algorithms don't favour images that show a lot of skin, the share of racy images from monitored accounts that data donors encounter should be roughly equal to the share of racy images monitored accounts post. If, however, data donors encounter such posts more frequently than their share in the posts created, that would support our hypothesis of the Instagram algorithm favouring these types of posts. The relevant metric for our analysis was therefore:

> `share in encounters / share in created posts`

This measures how much higher the share of encounters is than the share of created posts. If the value is above 1, an encounter is more likely than expected. If it is between 0 and 1, it is less likely.

We calculated this metric for each of our three indicators (raciness, female nudity and male nudity) across all available data donations. In a further step, we also conducted the same analysis for each individual data donor.

Statistical hypothesis testing

To see whether the differences found were statistically significant, we conducted a one sided two-proportions z-test (for significance level $\alpha = 0.05$). This method can be used to test whether the proportions in several groups are equal.

In our case, the null hypothesis H_0 to be tested was: the share of encounters with posts containing nudity is less than or equal to the share of created posts containing nudity.

$$H_0: \text{share}_{\text{nudity}}(\text{encounters}) \leq \text{share}_{\text{nudity}}(\text{created posts})$$
$$H_1: \text{share}_{\text{nudity}}(\text{encounters}) > \text{share}_{\text{nudity}}(\text{created posts})$$

If the test outputs a so-called p-value lower than the chosen significance level α (0.05 in this case), the null hypothesis is rejected in favour of the alternative hypothesis.

Statistical significance is an indication that differences observed in the sample groups are likely not a product of chance, but rather point to a real difference in the underlying population. The smaller the significance level, the more conservative the test is. The bigger the sample size and the larger the difference between the two groups, the more likely it is that the null hypothesis can be rejected.

Alternatives explored

We decided on the metric illustrated above as the most robust measure of algorithmic bias we could construct from the available data. During the course of our analysis, we explored the possibility of using other metrics to calculate the relative success of posts.

Share in encountered posts vs shares in created posts

Using the share in encounters as the numerator for our metric means that multiple encounters with the same post will be counted separately. A variation would have been to count the number of individual posts encountered instead:

```
share in unique encountered posts / share in created
posts
```

This metric would have explored the question of whether a post was likely to be shown to audiences at all, while counting each encounter also takes into account how often a post was shown. We decided to focus on the latter, since the repetition in donor's newsfeed is an indicator of algorithmic bias in itself. Repetition may also make it more likely that a user interacts with a post, contributing to its success.

Share of possible audience

We also considered another metric describing which share of its possible audience a post reached. The possible audience of a post would be the number of data donors who follow the creator's account. The actual audience would be the number of data donors who encountered the post at least once. This would be summed across all posts in the relevant category.

```
actual audience / potential audience =
# of donors who saw post / # of donors who follow account
```

If posts containing nudity reached a higher share of their potential audience than the average post, that would support our hypothesis. Like the approach described in the paragraph above, though, this metric would focus only on whether a post was shown at all, excluding the aspect of repetition. We also found it harder to interpret than the metric we decided on using, not least because the average shares were relatively small to begin with (the average post reached around 10-15% of its possible audience).

Results

For both raciness and nudity from either gender, we found that the share of encounters was significantly higher ($\alpha = 0.05$) than the share of posts created by the monitored accounts.

Raciness

705 of the 1831 image posts by monitored accounts with at least one donor follower contained images labelled "likely" or "very likely" racy by the Google Vision API. That amounts to 38.5% of created posts. But in encounters by data donors, the share was significantly higher: posts containing racy images made up 48.3% (1413 of 2924) of encounters.

This effect was observed both for posts by female and by male users. Posts by female users containing racy images were 23% more likely in encounters (51.7% vs. 42.1%). For men, they were 38% more likely (36.6% vs. 26.5%).

Nudity

We found that 13.3 % of posts (244 of 1831) contained pictures with labels indicating female nudity ("Brassiere", "Lingerie", "Undergarment" or "Bikini"). Among the encounters by data donors, that share was 56% higher: 609 of 2924 encounters (20.8%) contained female nudity.

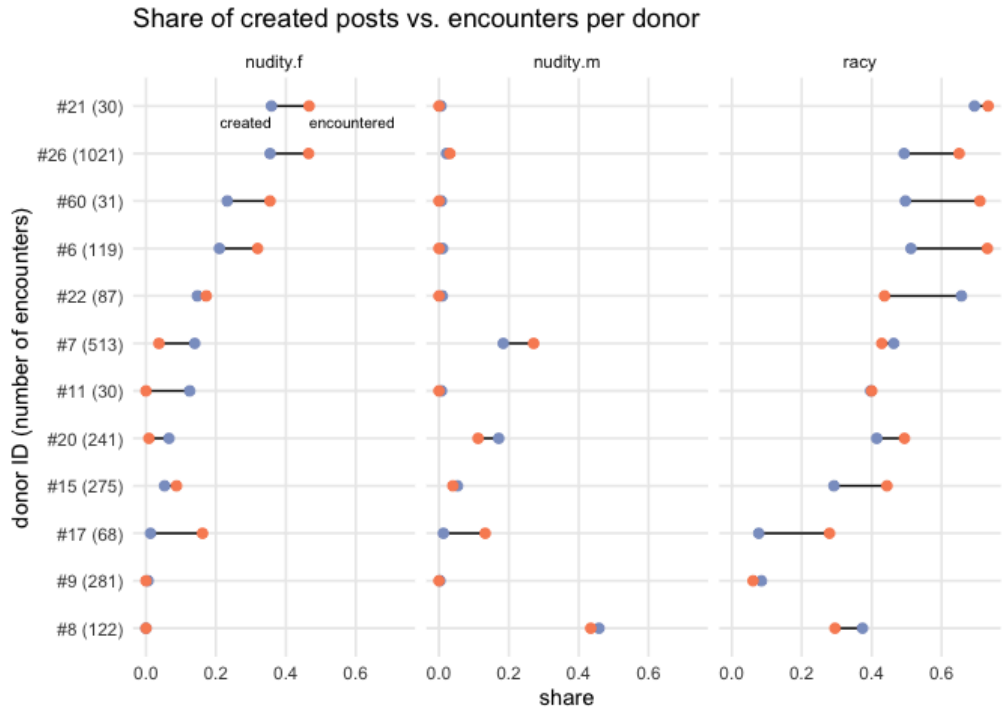
Male nudity was rarer: 129 of 1831 posts (7%) contained images with labels indicating male nudity ("Bodybuilder" or "Barechested"), while 9.5% of encounters did (278 of 2924). That makes male nudity 35% more likely in the feeds of data donors.

In both cases, the difference was found to be statistically significant.

Individual donors

Finding out how much the observed results differ by donor gives a good indication of how robust our analysis is, as well as of the degree of personalization Instagram's algorithm might apply.

We calculated the metric described above our three indicators for each donor with more than 30 donations, comparing the share in their encounters with the share of posts by the monitored accounts they follow. While the majority of donors exhibited the same patterns observed in the overall trend, some (around 1/3) did not.



Notably, wherever there was a large difference between the share of encounters and the share in created posts, it was usually in favour of a higher share in encounters.

These observations would be consistent with a situation where Instagram's algorithm favours posts with nudity by default, but also takes into account specific user preference, which would dampen the effect in some cases and amplify it in others.

Limitations

With 31 data donors total, 26 of which contributed a donation at least once, our analysis necessarily depends on donor behaviour to a certain degree.

At the moment, a large share of data donations come from very few donors, making the results of our analysis vulnerable to the specific behaviour and preferences of these donors. They might, for example, have disproportionately followed accounts which post a lot of racy content, or have a strong personal preference for or against nudity. We have taken this into

account by including an analysis for each individual donor but can't rule out such effects altogether.

Our analysis also works under the assumption that each data donor could, theoretically, have encountered each of the posts by the monitored accounts they follow. In reality, this depends on the timing and frequency of posts and data donations.

To better study our hypothesis, a larger sample size would be needed, both in the number of monitored accounts and in the number of data donors. This would allow us to better mitigate the effects of differences in behaviour and draw more reliable conclusions. So far, our observation has been that the results presented above become more robust the more data is gathered.

Summary

Our analysis presents strong evidence that pictures which show more skin are shown to users more often than pictures that don't. Sexually suggestive images, as well as nudity from either gender, appeared significantly more often in data donors' newsfeeds than in the posts created by monitored accounts. This effect was observed in most, if not all, of our data donors. Larger sample sizes would be needed to further consolidate our results.