Assignment #4

Gaurav Lalwani
Sonakshi Garg
Shweta Mane
Milind Siddhanti
Sanchit Singhal
Pedro Seguel

Fall 2018, November 19, 2018

MIS 381N – User Generated Content Analytics w/ Dr. Anitesh Barua

McCombs School of Business, the University of Texas at Austin

B) What accuracy do you get by using the post_caption words as the independent variables instead of image_labels? Finally, what accuracy do you get by combining (concatenating) the image_labels and post_caption and using them together as independent variables? What can you conclude from your analysis?

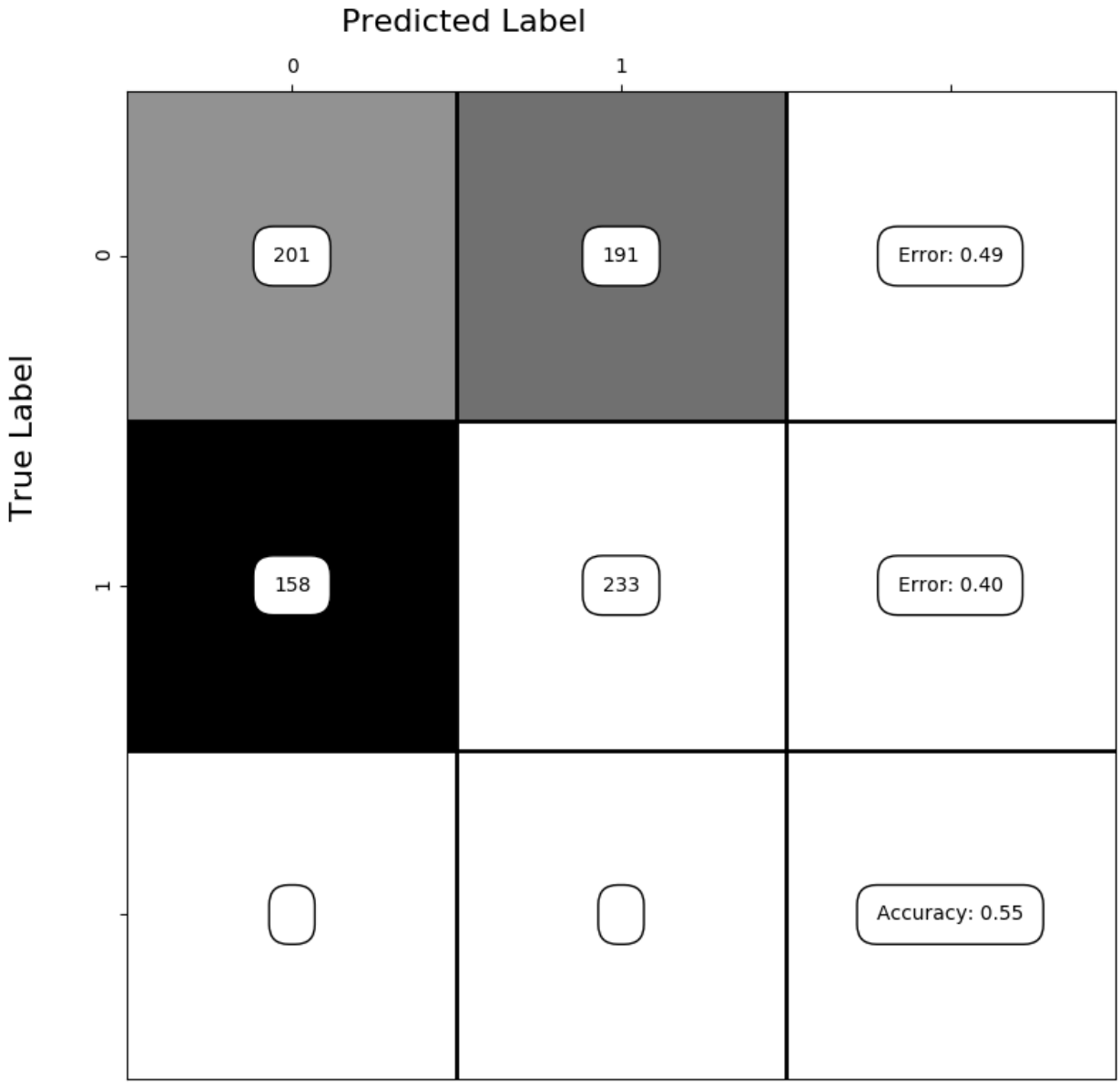

Figure 1: Image labels as the independent variable

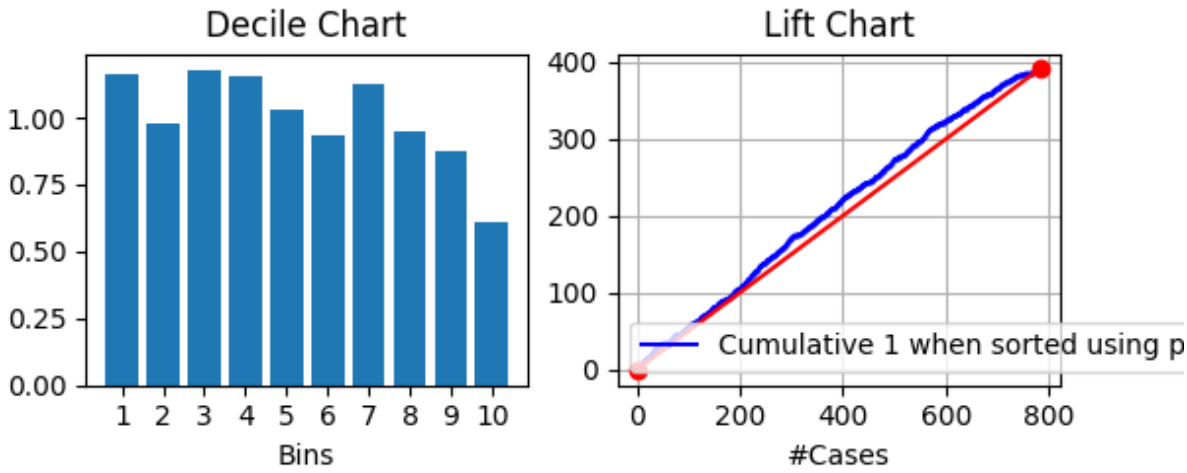

Figure 2: Graph displaying lift values with Image Labels as the independent variable

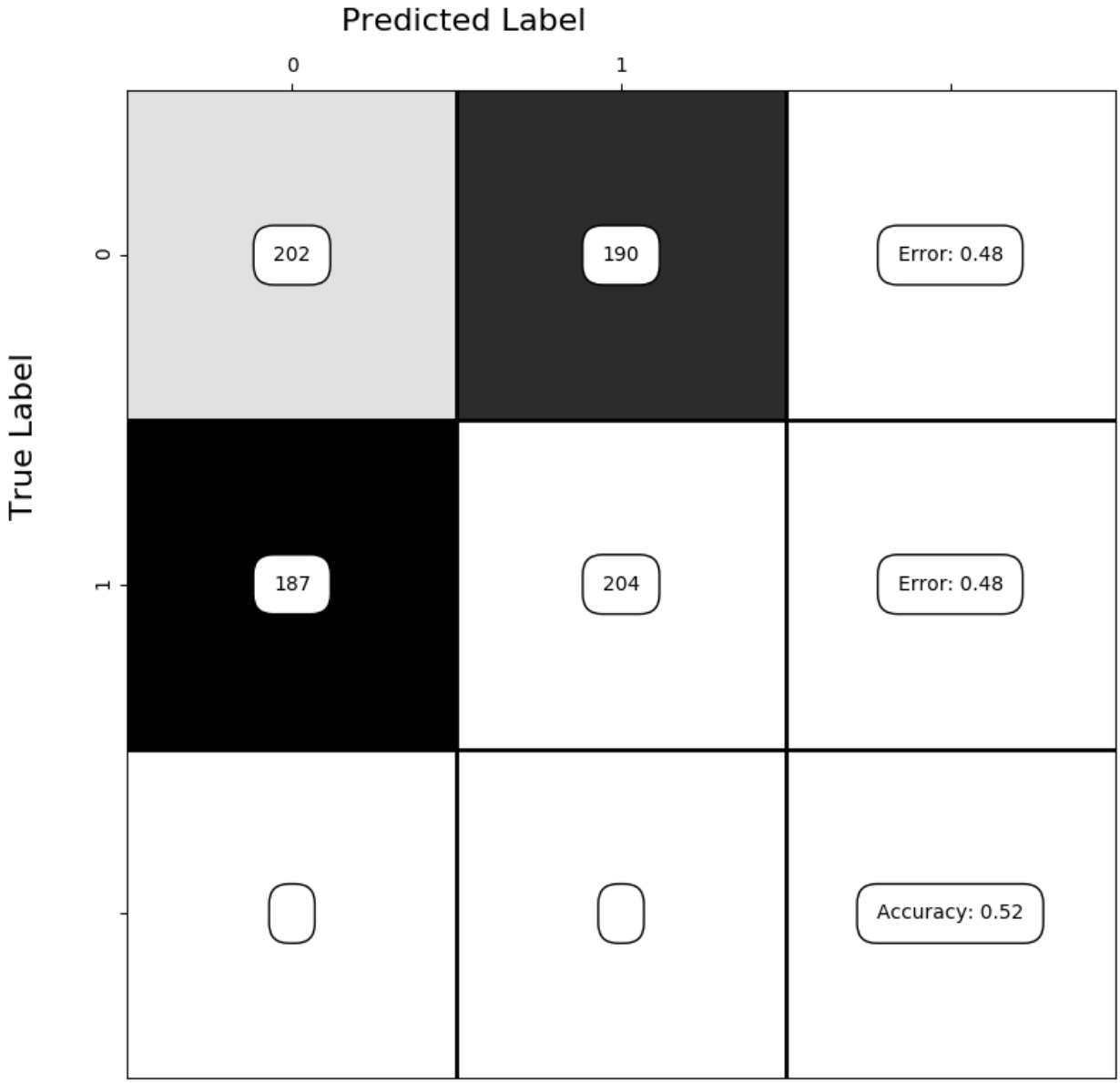

Figure 2: Post Captions as the independent variable

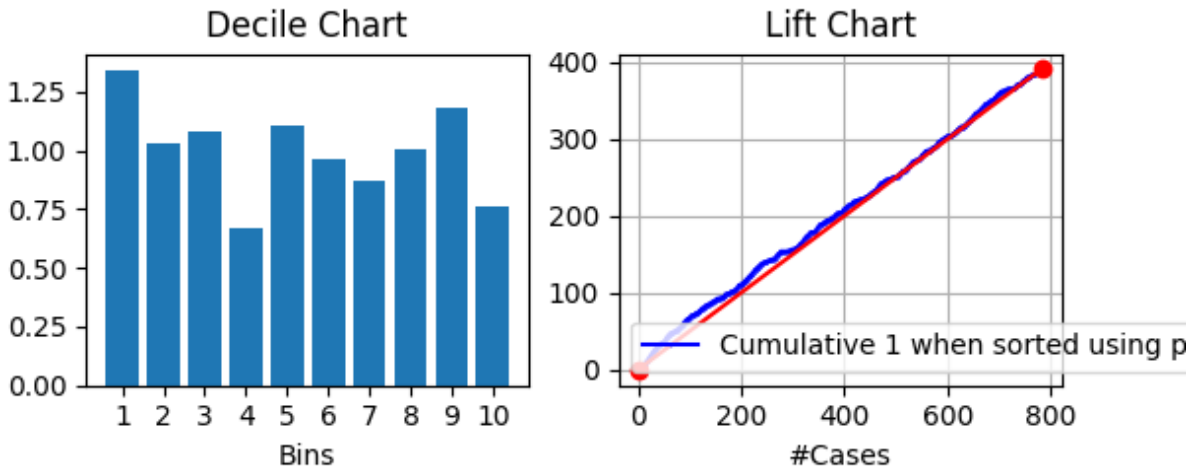

Figure: Graph displaying lift values with Post Captions as the independent variable

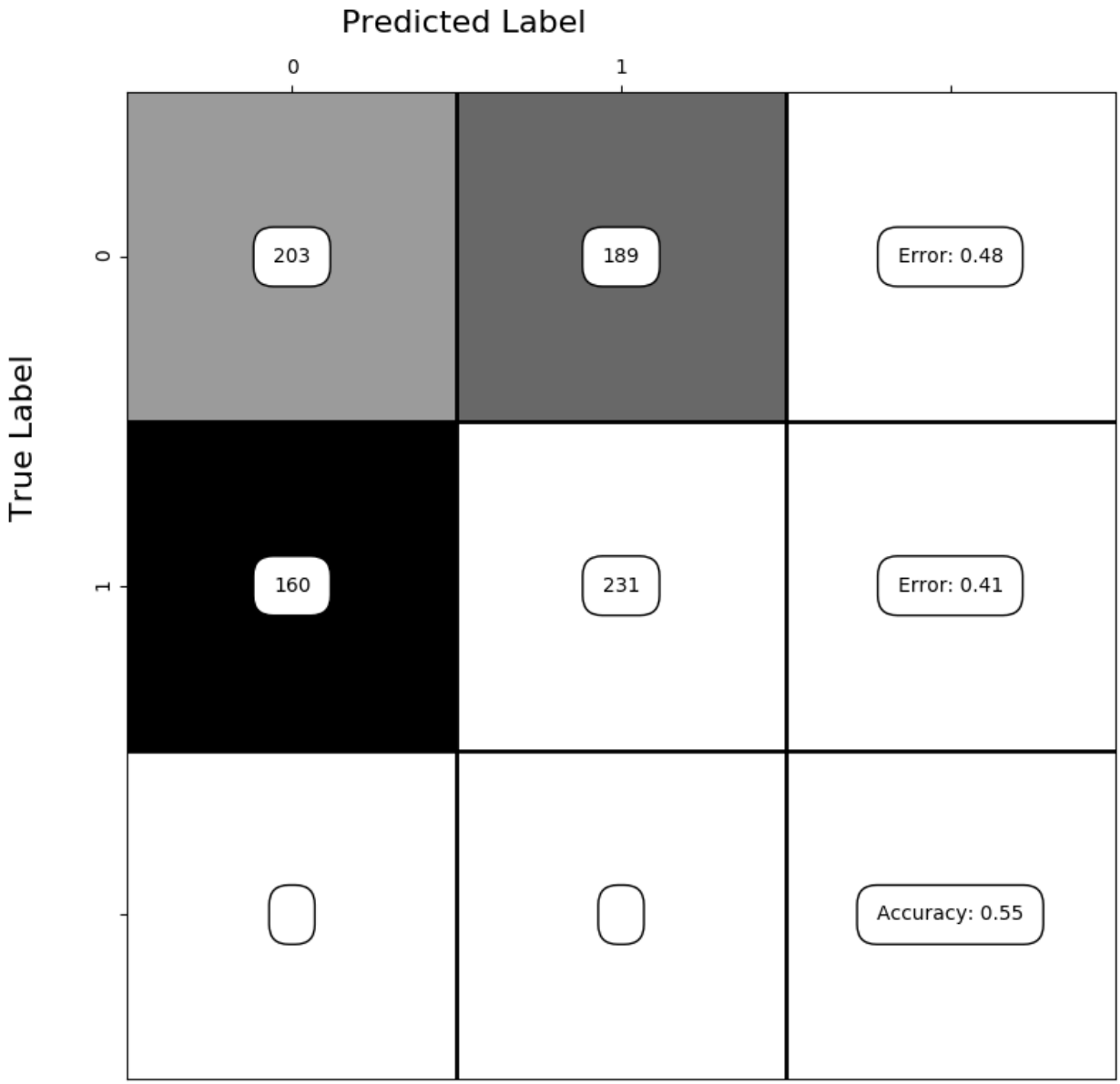

Figure 3: Combination of Image Labels and Post Captions as the Independent Variable

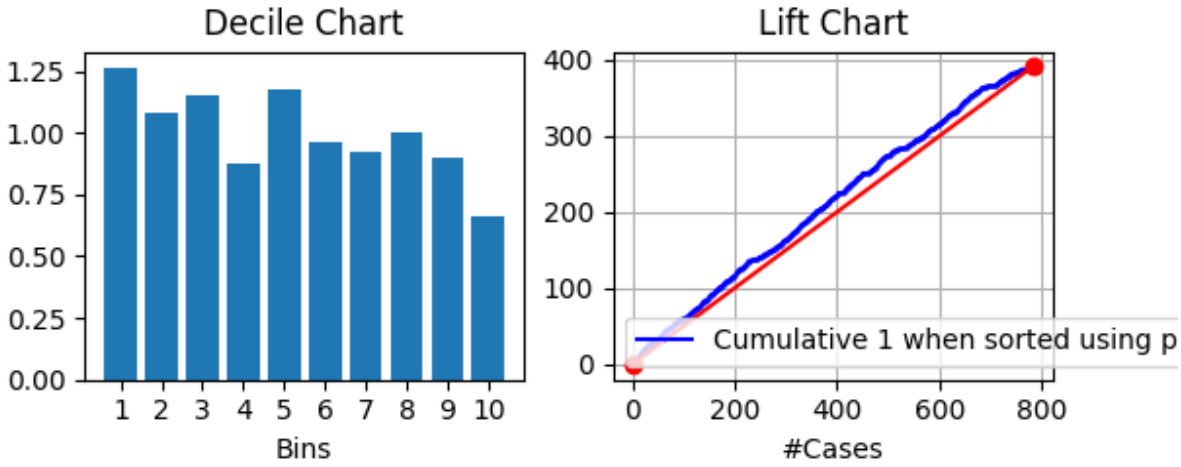

Figure: Graph displaying lift values with combination of image values and post captions as the independent variable

From above analysis, we can see that accuracy for Image labels is higher (but only slightly) than that of post captions when calculated independently. Combining image labels and post captions did not improve the accuracy much either. The reason for a lower score when combining the images and post captions might be multicollinearity and therefore it leads to a lower overall accuracy. It is also possible that insufficient text on Zara's instagram posts does not provide much insight into engagement levels of the posts and therefore does not provide a significant increase in our model's accuracy over and above the image labels. This means that on a platform like Instagram, the images are leading to higher engagement levels than the text on the captions - at least on Zara's account.

C) i) Topic modeling

We conducted a topic modeling analysis, finding 4 main topics: 1) Shoes & Accessories 2) Photography 3) Outwear clothing 4) Overall design. We chose these names using the image labels with the highest scores in the LDA. Table 1 provides an example of our rationale.

It should be noticed that we iterate this process with 5, 2, and 3 topics. We found 4 to be the cleanest match of image labels grouped in specific topics.

Table 1. Topic modeling based on Image Labels (showing 10 labels with higher scores for each topic)							
Shoes & Accessories		**Photography**		**Outwear clothing**		**Overall design**	
Image_labels	**0**	**Image_labels**	**1**	**Image_labels**	**2**	**Image_labels**	**3**
product	0.11498722	**photography**	0.07710764	**outerwear**	0.09361183	**fashion**	0.1259148
shoe	0.06468381	**hair**	0.06899158	**fashion**	0.07021012	**model**	0.08034364
design	0.04232673	**black**	0.06036827	**coat**	0.06025194	**shoulder**	0.0722421
footwear	0.03993133	**monochrome**	0.04717967	**neck**	0.05328122	**neck**	0.06515325
font	0.03274513	**model**	0.04667242	**model**	0.05128958	**clothing**	0.06110248
leg	0.03194666	**girl**	0.0446434	**sleeve**	0.04730631	**sleeve**	0.05975223
bag	0.02795433	**white**	0.04413615	**formal**	0.04431886	**joint**	0.05367607
shoulder	0.02715586	**shoot**	0.03399107	**wear**	0.04431886	**dress**	0.05030043
brand	0.02476046	**photo**	0.03196206	**jacket**	0.04232723	**pattern**	0.04388671
girl	0.02236506	**shoulder**	0.02942579	**clothing**	0.04033559	**outerwear**	0.03308466

ii) Table 2 shows the average topic weights by the lowest and highest quartiles. We got the quartiles by sorting the data from high to low engagement score and then calculating average topic weights for each topic.

There are observable differences between quartiles averages at the first (Shoes & Accessories) and third (Outwear clothing) topics. In Shoes & Accessories, the fourth quartile (0.29084291) has a bigger average weight than the first quartile (0.14602209). On the contrary, in Outwear clothing, the first quartile (0.32536447) has a bigger average weight than the fourth quartile (0.17568168).

We considered that the differences between quartiles below 0.1 were not relatively significant to consider.

Table 2. Average topic weights by the lowest and the highest quartiles

	Shoes & Accessories	Photography	Outwear clothing	Overall design
1st Quartile of engagement lowest scores	0.14602209	0.22621153	0.32536447	0.3024019
4th Quartile of engagement highest scores	0.29084291	0.21892316	0.17568168	0.31455225
Difference	+0.144820824	-0.007288375	-0.14968279	+0.012150341

D) What advice would you give Zara if it wants to increase engagement on its Instagram page based on your findings in Tasks B and C?

Based on our analysis in part B, it is evident that the images on Zara's instagram posts are a better indicator of engagement as compared to the captions - even a combination of the two did provide additional predictiveness of engagement. Therefore, we would advise Zara to focus more on the images that being selected as opposed to its caption in order to increase engagement levels.

From our analysis in part C, it can also be seen that the largest difference in the average topic weights of images across two quartiles is in Outwear clothing and Shoes/Accessories. This tells us that posts related to these two topics show a significant variance in engagement. In contrast, the topics of photography and overall design do not show a difference in topic weights across the highest and lowest quartiles suggesting that users are not displaying much variance in engagement for these topics and hence are not good predictors of engagement scores. When investigating Outwear clothing more closely, it can be observed that average topic weight is actually lower in the upper quartile of engagement than the lower quartile. This means that Outwear Clothing is in fact leading to a lower level of engagement and Zara should refrain from posting about it. Shoes/Accessories are the opposite, posts related to this topic lead to higher

engagement levels; the upper quartile of engagement scores has a significantly larger weight of Shoes/Accessories. Although we have mentioned that Overall Design topic does not show much variance, it is still the highest weight topic score in the upper quartile of engagement.

Putting these considerations together, we would advise Zara to focus their marketing efforts more on the images that are being selected than the caption in order to increase engagement on its Instagram page. It is also advised that the images be around the topic of Shoes and Accessories as that has proven to lead to higher engagement. Overall Design is another topic that, although does not increase levels, but has displayed a significant average weight in high engagement posts and Zara should continue to maintain these images. We would also advice Zara to reduce images regarding Outerwear clothing as this as proven to explicitly decrease engagement levels on their instagram page.