DSCI 550
04/6/2023
Professor Mattmann
Assignment 2
Team 12: Jai Agrawal, Daniil Abbruzzese, Todd Gavin, Tania Dawood

Pixstory Extracted Insights

For Assignment 2, we built on our assignment 1 work to explore the unexplored areas of the Pixstory dataset, with a focus on extracting large-scale content and implementing data science techniques. Our objective was to investigate if Pixstory users adhere to the stated goals and intended use of being a “clean” social media platform. In order for us to explore our objective, we began by using Google Language Detect and Tika Language Detect to identify the standard two character language identification codes for the Pixstory dataset posts using the “Narrative” and “Title” columns. Once we had the languages identified, we were able to use RTG and Tika translate to automatically generate translations of all text in the Narrative and Title columns. Our next step was to use GeoTopicParser and geocode to identify the locations associated with the posts. Then, we used Detoxify to classify the toxicity of the posts and compare the Detoxify score to the ADL, GLAAD and sarcasms flags we generated in assignment 1. Lastly, we leveraged the richness in the underlying images associated with the Pixstory posts to generate a caption as well as a list of objects present in the imagery. These additional features allowed us to examine the posts and compare if the images, captions and objects would uphold the stated goals of the platform. Additionally, we also analyzed the data to identify any trends in the image analysis with respect to age, gender or posting dates and time.

Note: The completed dataset (Master_Pixstory_Dataset_Complete.tsv) can be found inside the repository with path: */DSCI550-PixstoryMediaExtractionAndAnalysis/Datasets/Master_Pixstory_EXTRACT_Dataset_Complete.tsv*

Key insights

As mentioned above, our first step was to use Tika and Google Lang detect on our dataset, to detect the language codes. Once we completed this, we found that the most frequent language for posts was English and the least prevalent language was Telugu. However, one thing we noticed in the results was that Tika and Google Lang Detect had some differences in identifying the languages. For example, Tika identified 67056 narrative posts as English whereas Google identified 67663 posts as English. These differences were also prevalent for a majority of the languages as can be seen incFigure 1 below. Another thing we noticed was that on occasion the “Title” language detection would not match the “Narrative” language detected. We saw this issue with both Tika and Google. For this reason, we would need to do a more indepth language analysis to confirm our insights. (Question 2)

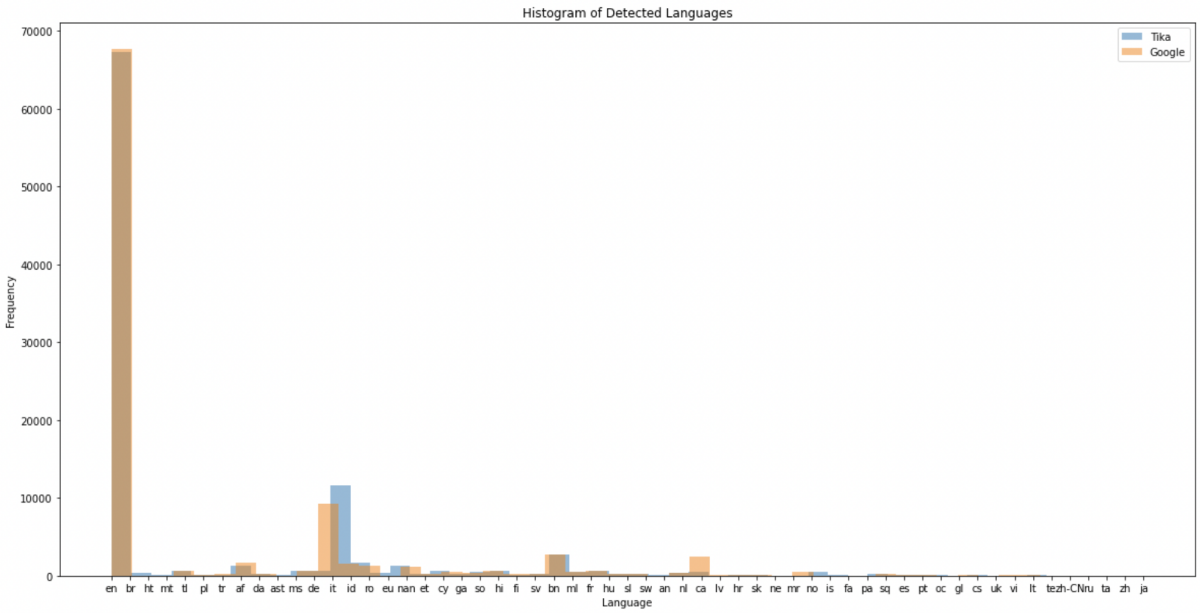


Figure 1: Narrative Language Detection Using Tika and Google Lang Detect

Once we were able to identify the language of each post, we could begin our work on using the location to discover trends in the data relevant to age, gender and interests. To do this, we created a heatmaps as seen in Figures 2-4 below.

Age:

The heatmap in figure 2 shows the relationship between Age (Y axis) and GeoTopic Name (location) (X axis) through which we can see several significant correlations can be identified. Based on this, we can see that the GeoTopic name of "British Indian Ocean Territory" has the greatest clustering of associating user narratives with ages between 18 and 24, with a significant number of users between the ages of 42 and 47. We believe that this is accurate because Pixstory is a widely used social media platform in India. Additionally, the GeoTopic Name "Islamic Republic of Afghanistan" has a cluster of associated user narratives with ages between 21 and 22, while the GeoTopic Name "Manchester" has a cluster of associated user narratives with ages between 21 and 24. Overall, the highest concentration of associated user narratives for all GeoTopic Names was between 18 and 26.

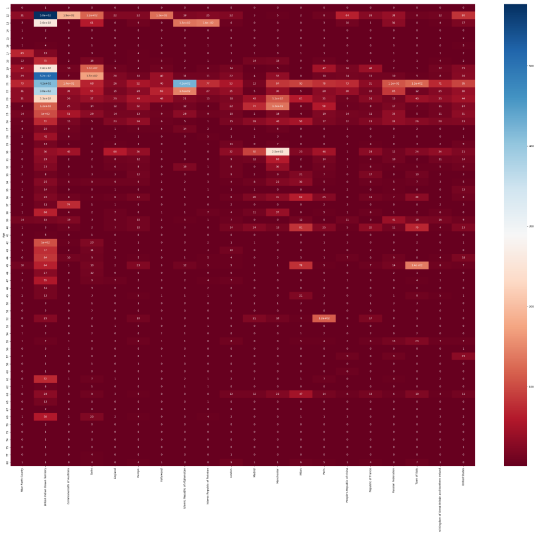


Figure 2: Age Correlated with GeoTopic Location

Gender:

The heatmap in figure 3 shows the relationship between Gender (Y axis) and GeoTopic Name (location) (X axis). The heatmap shows that the GeoTopic name of "British Indian Ocean Territory" has the greatest number of associating user narratives for males over females. On the other hand, the "Islamic Republic of Afghanistan" is the only GeoTopic with more females than males.

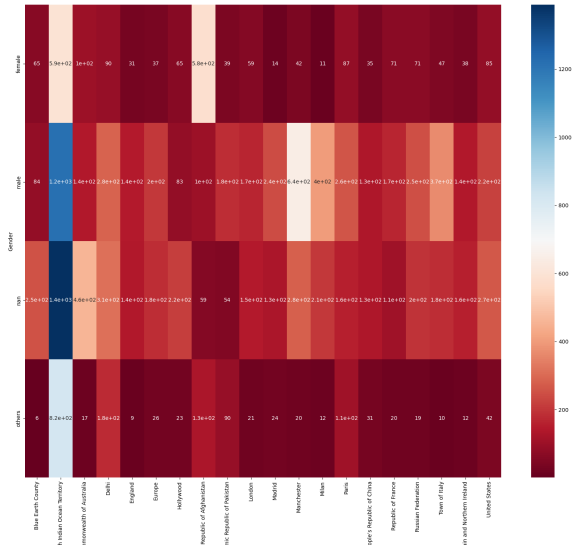


Figure 3: Gender Correlated with GeoTopic Location

Interests:

Finally the heatmap in figure 4 shows the relationship between Interests (Y axis) and GeoTopic Name (location) (X axis). Through this heatmap we can observe that the GeoTopic name of "British Indian Ocean Territory" has the greatest greatest clustering for multiple interests such as health, politics, and sports.However, the GeoTopic "Russian Federation", shows a high amount if Interest clustering for topics surrounding the Russia-Ukraine War such as war, Ukraine war, Ukraine, US Government, Vladimir Putin and others.

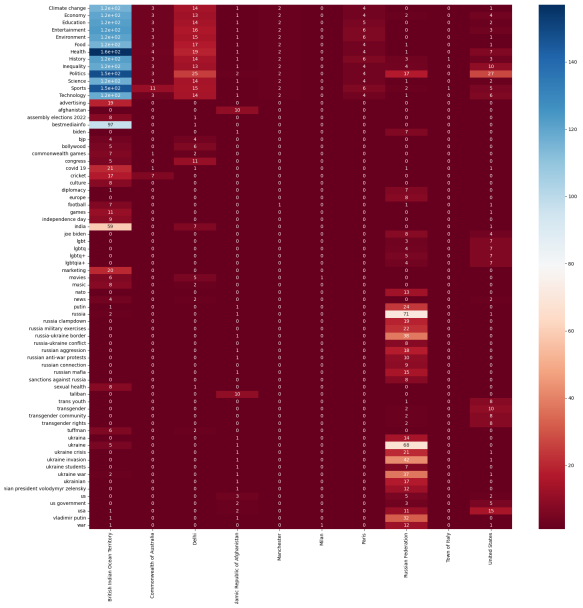


Figure 4: Interests Correlated with GeoTopic Location

Overall, insights from the heatmap are valuable for understanding the different age and gender demographics, in addition to general interests of users in various locations. This information can be used for further research. (Question 1)

Now taking into consideration the location and language features, we created a heatmap for language detection using Tika and Google Lang Detect (Figure 5). We can see that a majority of the clustering is for language code en (English). The largest clusterings though, are for the GeoTopics of "British Indian Ocean Territory", "Delhi", "Manchester", and "Islamic Republic of Afghanistan". Behind the language code en (english), the second highest language code is it (italian), with of course the GeoTopic "Town of Italy" (and other related GeoTopics) have the highest clustering. (Question 3)

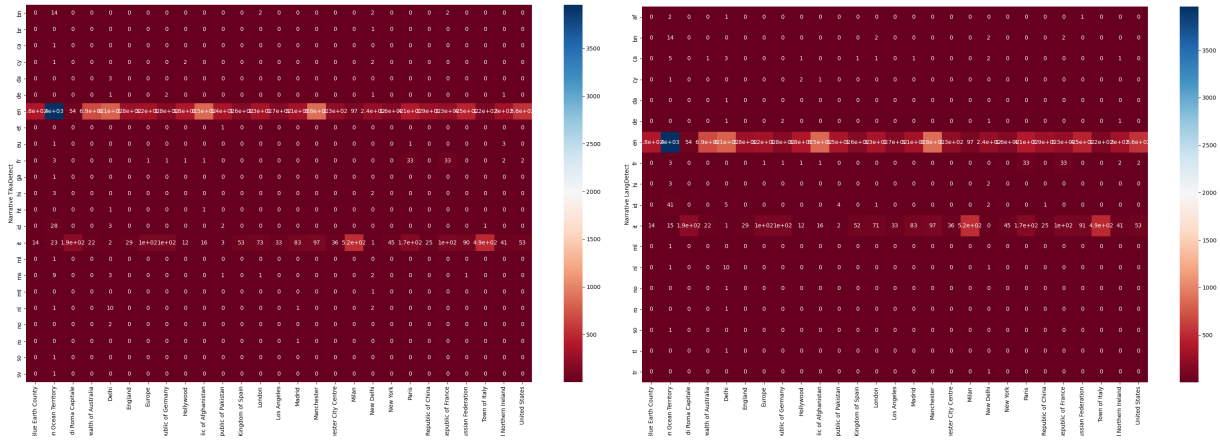


Figure 5: Tika (Left) and Google (Right) Language Detection Distribution

Digging deeper into the location of posts and entertainment and sporting events, we wanted to check if there was any correlation between the three features. Initially we sampled our dataset twenty times to identify the correlation between locations but we found no correlation between location of posting and event. On the other hand, when we looked at the top talked about entertainment events and sporting events, as well as the top posting locations, we were able to derive some insights. We noticed that Australian events were some of the most talked about events and one of the most posted from locations was also within Australia. We also discovered that the most posted about sporting events included the Africa Cup of Nations and the Australian Open, while the most talked about film festivals were the Melbourne International Film Festival and Sarajevo Film Festival and the Edinburgh International Film Festival. Additionally, the locations that were most frequently posted from were the British Indian Ocean Territory, Manchester, Delhi, the Islamic Republic of Afghanistan and the Commonwealth of Australia. However, a deeper analysis into location of posting and events could provide us with a clearer picture of whether correlations between post location and events exist. (Question 4)

Next we classified the posts using Detoxify and compared each post's score to its GLAAD, ADL Flags. We findings showed that there is not much correlation between hate speech, sarcasm flags and detoxify scores as shown in Figure 7. We believe this could be due to a faulty score generation or it could simply be that there is not much of a relationship between them. However, our analysis indicated a strong correlation between obscene and insult posts with toxicity. Which is something we expected as the two are related. The heat map also shows that obscene posts can sometimes be associated with severe toxicity. Overall, our detoxify analysis suggests that different types of posts toxic content might be correlated with each other, but not necessarily with other types of negative language and hate speech or sarcasm. (Question 5)

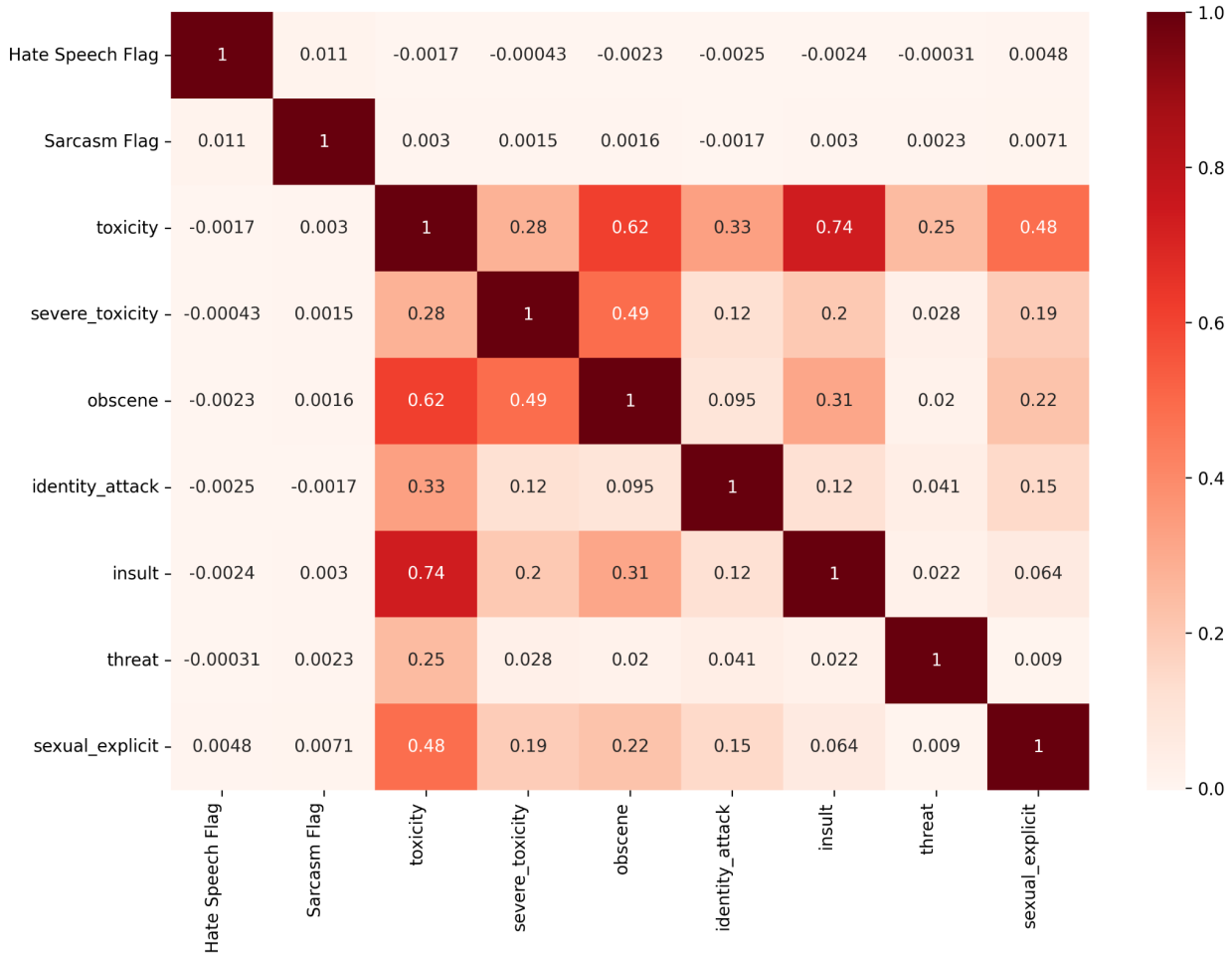


Figure 7: Pearson Correlation Scores for the GLAAD, ADL Flags & Detoxify Scores

After implementing the caption generation model, we found that the accuracy of image captions can be variable. Sometimes the model can provide an accurate description of the image, but other times it gives a completely incorrect image description. However, we found that the model is good at recognizing certain elements of the image including colors, people, their gender and certain objects. This model is also capable of accurately describing the positioning of different elements within the photo that are positioned relative to each other. For example, it can identify object X is lying in front of object Y.

On the other hand, the model is not very good at recognizing the location of the images. Instead of directly recognizing the location, it makes inferences based on the elements present in the photo. For example in the photo shown in Figure 8 the model identified there is a lot of white color in the photo and deduced that the subject was in snow when it was actually engulfed in fog.

Figure 8: (Pixstory-image-164441526830771.jpeg)

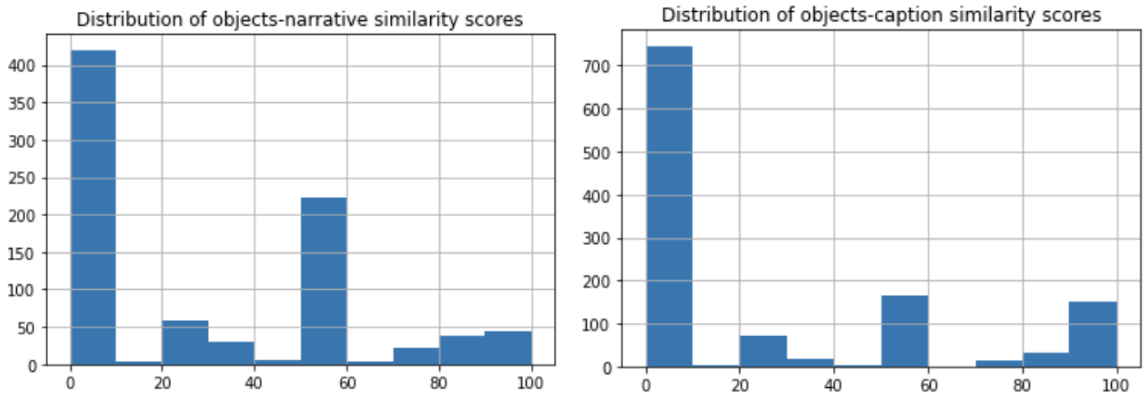
Additionally, it may not be great at deducing the actions of people in the image, especially when there are no contextual objects present to represent the situation. (Question 6)

To get more granular in our analysis, In the python file “q7_objects_in_narrative.ipyb” we attempted to run an analysis that examines if the identified objects from the image were described in the narrative of the post as well as if these objects were described in the generated caption. To do this, we used Open AI’s API to generate a score of relation between these features.

The reasoning behind this is that Open AI has a very sophisticated grasp on understanding language from text and the meaning behind it, and thus is able to easily identify synonyms of keywords. For example, we would expect it to understand that the word ‘basketball’ is highly related to “NBA," whereas a simple keyword analysis would detect these two as entirely different. Furthermore, Open AI’s model is sophisticated enough to find full sentences that relate to keywords, even if each word in on its own does not. For example, we would expect it to understand that the statement “he is one of the best defenders in the league” highly relates to “basketball,” even though each word on its own does not clearly indicate the discussion is about basketball.

We gave the api two prompts: promt 1 “I want you to output only a number between 0 and 100. This number represents how if these objects '{object_recognition}' relate to what is discussed in this narrative '{narrative}', where 0 means there is no relation and 100 means there is a high relation. Do not output any other text, only output just the number," and prompt 2 which is exactly the same but replaces ‘{narrative}’ with ‘{caption.}’ We generated these similarity scores for a sample of 1,200 different posts and calculated the mean for each. We found that narratives had a similarity score of 26.8 on average and captions had a similarity score of 23.5 on average. This indicates that the identified objects present in the image are not strongly related to what was described in the original post or the generated caption. Of the sample of 12,000, we found that there is an irregular distribution for both object-narrative similarity and object-caption similarity with the highest frequency of scores being in the 0-10, 50-60, and 90-10 range. This suggests inconsistency within the model, which likely is the result of being able to handle some types of photos very well and others very poorly. Furthermore, all of these conclusions are under

the assumption that OpenAI's API was accurate in its analysis of similarity, but more work needs to be done in the future to assess if it did a good job at following the prompt. (Question 7)



Lastly, we wanted to look into any age, or gender specific trends we could see in the text captions or identified objects in the image media. In order to generate these insights we looked at the most common captions, the mean age of users who posted an image that generated said caption as well as the gender distribution. We followed the same process for identified objects. The results from the top three captions and identified objects are as follows.

Captions:

1. "a group of people standing next to each other" (5022), the mean age is 25.95, and the gender distribution is 60.4% male, 26.4% female and 13.1% others (approx).
2. "a group of men standing next to each other" (2118), the mean age is 25.84, and the gender distribution is 60.9% male, 26.8% female and 12% others (approx). (similar to the prev one)
3. "a group of young men playing a game of soccer" (1438), the mean age is 26.08, and the gender distribution is 62.5% male, 24.1% female and 13.3% others (approx).

Identified Objects:

1. Website/internet site/site objects (2171), the mean age is 26.22, and the gender distribution is 65.3% male, 18.9% female and 15.6% others.
2. Comic book/book/dust cover objects (1151), the mean age is 25.67, and the gender distribution is 60.4% male, 26.7% female and 12.8% others.
3. Suit/clothes/tie objects (1116), the mean age is 25.62, and the gender distribution is 63.5% male, 23.4% female and 12.9% others.

Overall, for image captions, we noticed the trend that the mean age is somewhere between 25-26 and the general gender distribution is skewed towards men (~60%) while females are around 24-26%. The demographics did not change much for specific objects, as the mean age remained around 25-26, and the gender distribution is similar. However, an interesting discovery we found was that the distribution for females drops when the website object comes up. One explanation for this drop could be that women have different preferences and interests when it comes to the type of content they share online. Women are more likely to share content related to family, friends, personal interests whereas men are more likely to post about news, sports and current events which could more likely feature websites than a woman's content. (Question 8)

Bottlenecks

We did run into some issues while trying to conduct this analysis. For starters, while trying to run translations using Tika and RTG, we constantly encountered 500 warnings which meant we were unable to establish a connection to the Tika server. The only way we could resolve this issue was by killing all java processes and then uninstalling and reinstalling Tika. Additionally, we also faced some initial difficulty with setting up the image links to be processed when running the Tika Image Captioning. However, once we set up a local port associated with the IP address for en0, we were able to use the software with ease. One thing we did notice though, was that the Tika Image Captioning was highly inconsistent with extremely accurate captions generated and times and completely unrelated captions at others. Similarly, the software required to set up GeoTopic Parser required a significant effort to set up and configure. However, once we were able to get it set up, it was reliable and easy to use as the implementation of the parser code is simple and straightforward and it delivers high quality results.

On the other hand, Detoxify was a quick and simple process for our group and it was straightforward to use. However, the effectiveness of this software is vague and still requires more analysis to be adequately assessed.

Conclusion

In conclusion, through language detection, English was found to be the most prevalent language for Pixstory posts and our analysis also relieved differences between the two methods used, Tika and Google Lang Detect. The age and gender analysis showed the British Indian Ocean Territory had the highest concentration of user posts between the ages of 18 and 24, and the greatest number of posts from males compared to females, and interests varied by location. Furthermore, Detoxify found a strong correlation between obscene and insult posts and toxicity, the caption generation models accuracy of image captions can be variable and the identified objects present in the image are not strongly related to what was described in the original post or the generated caption. Overall, we would recommend a continued and deeper analysis into Pixstory data to determine whether users are appropriately using the platform or not.