Report – Assignment 3 Team 3

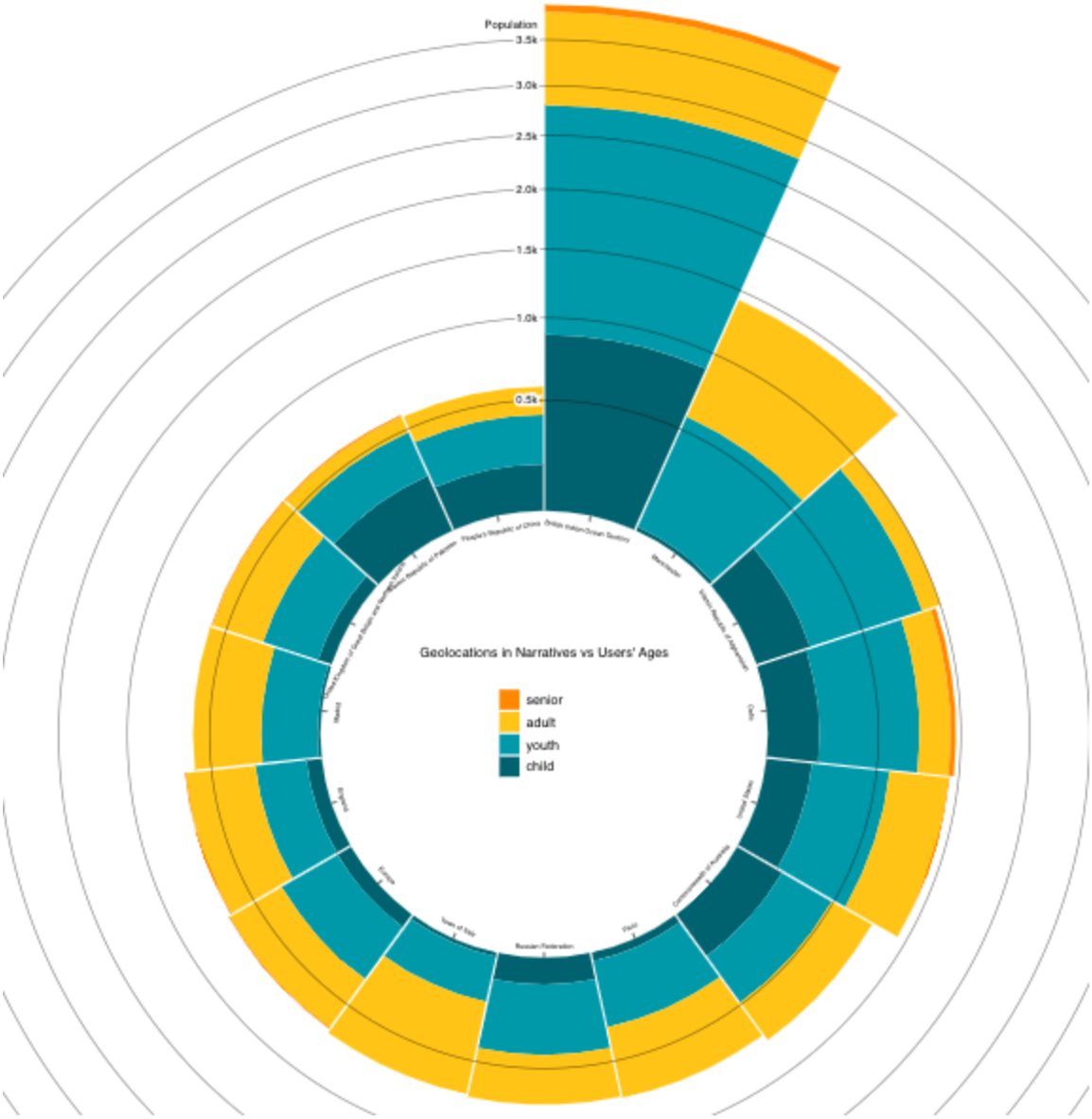
Jimin Ding
Xiaoyu Dong
Hui Qi
Mingyu Zong

1. Why did you select your 5 D3 visualizations? How are they answering and showing off your features from assignments 1 and 2 and the work you did?

Radial Stacked Bar Chart

Analyzing user groups and the geolocations mentioned by them can help the PixStory Application better understand how users of different ages pay more attention to things happening in specific places such that it could recommend posts better to fit their interests and concerns.

We use a radial stacked bar chart to show the relationship between the places mentioned in the narratives and the user's age distribution. We also did a similar analysis and visualization in assignment 2 with a regular bar chart. The following radial stacked bar chart can highlight the trend and outlier in the data. For instance, British Indian Ocean Territory in the figure below is mentioned more than 3,500 times, while most locations are between 500 and 1,000 times. Senior users mainly wrote posts about "British Indian Ocean Territory" and "Delhi." Moreover, adult users mentioned the Islamic Republic of Pakistan and the Islamic Republic of Afghanistan fewer than other places



Index Chart

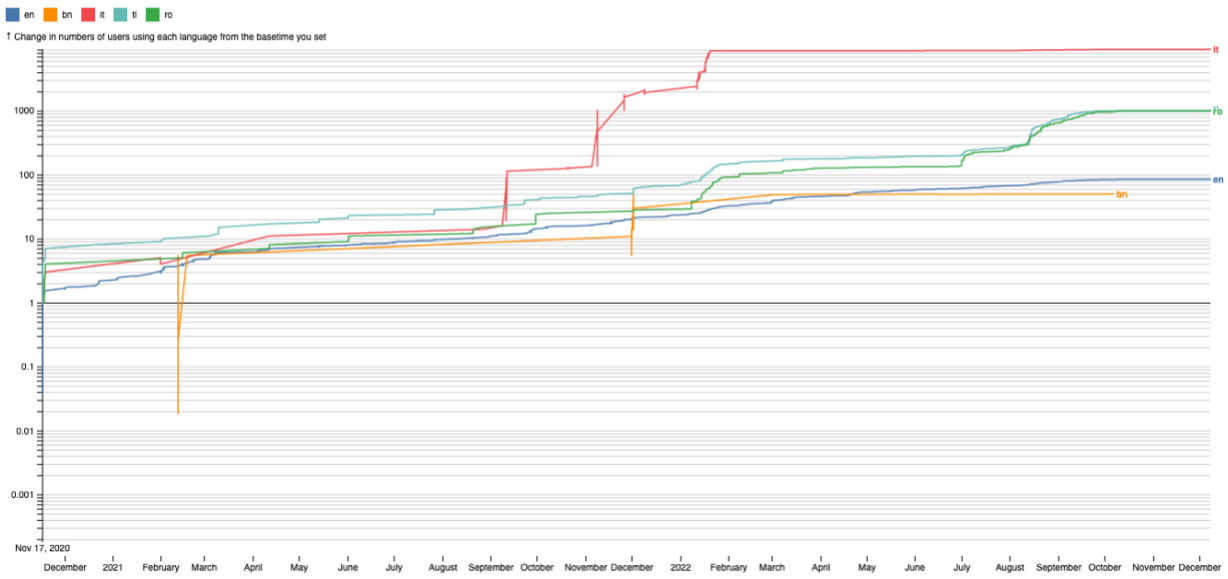
In Assignment 2, we specialize in distinguishing the language of posts because contrasting user language groups can help us better understand the distribution and needs of users who use this software. We can also formulate some business strategies such that users would be more willing to introduce the PixStory application to their friends with this software. The index chart below shows the growth in the number of posts in the top 5 languages over the past two years. By moving the vertical line to change the baseline time, we can compare the growth of usage of each language from a specific time base to the end of 2022 on these existing past data.

If we set November 2020 as the base time, by the end of 2022, "it (Italian)" will have the most significant overall growth. "Ro (Romanian)" and "tl (Tagalog)" will have overall similar growth. In contrast, "bn (Bengali)" is the least, and English's development is the second least. But, English (en) is already the largest user group at the base time. Thus, although the total number of posts written in English has not grown much, the total number of users using English is still the largest.

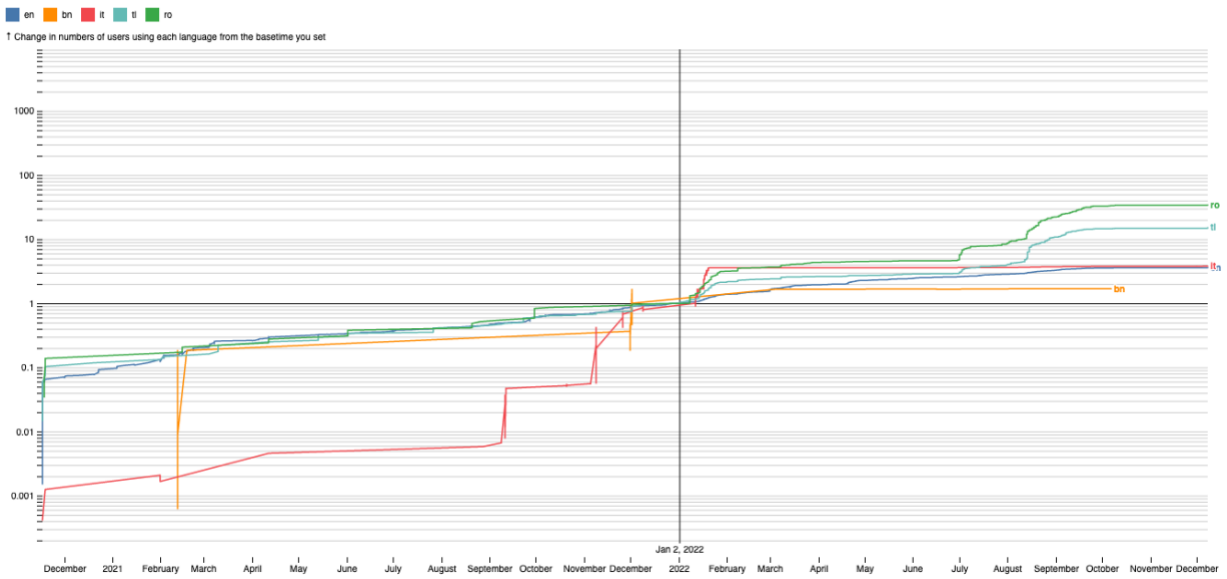
In addition, in September 2021, "it (Italian)" growth speed exceeded "tl (Tagalog)," which had the most considerable growth previously. What's more, from the figure below, the lines of each language at the end of the graph are almost parallel to the x-axis, which implies each language has a few increases.

Index Chart

(Uses' languages in their posts through time)



When the base time becomes January 2022, "ro (Romanian)" grows the most, then "tl (Tagalog)" and "en (English)" grows almost the same, while "bn (Bengali)" is the least. From the figure below, "it (Italian)" had increased from January 2022 to February 2022, but after that, the line became flat, indicating that users who used Italian later had few increases. Furthermore, these five languages experienced a few increases until July 2022. After that, "ro" and "tl" began to grow again, and other languages still had little growth.

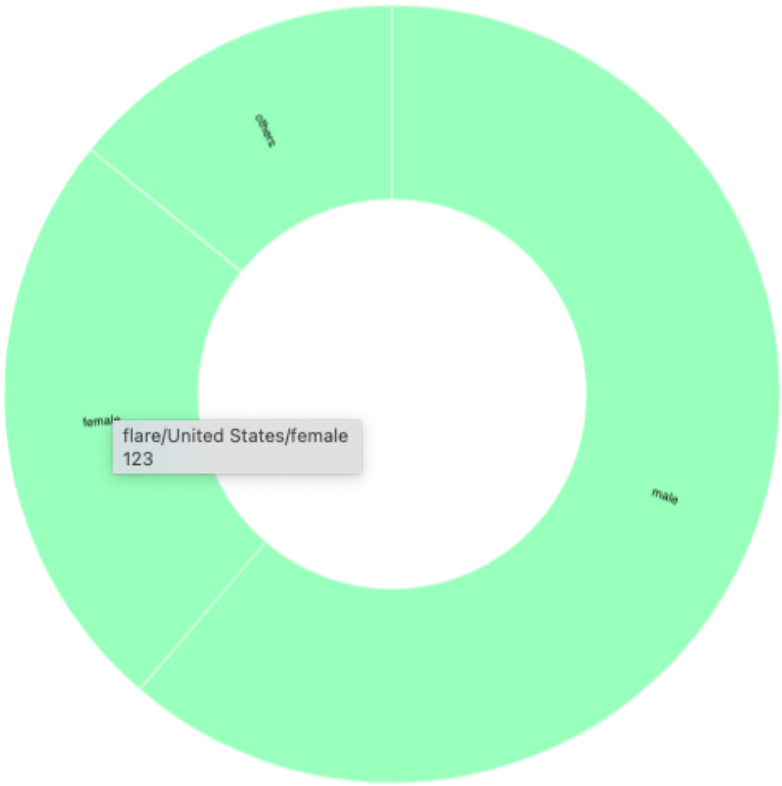


Zoomable Sunburst

In Assignment 2, we studied the location mentioned in narratives and user information, such as gender, age, language, and interests. In this assignment, we picked the ten places mentioned most in posts and used Zoomable Sunburst to see the distribution and proportion of users' languages and genders more intuitively.

The figure below shows the proportion of geolocation mentioned within the top ten geolocations in users' posts. At the same time, each geolocation provides the gender distribution of users who wrote those locations in their narratives. Places like "Madrid," "Russian Federation," "Town of Italy," and "Manchester" are mentioned mainly by male users rather than that by females and other genders. Male users generally wrote locations more than females and others, which matches the bar charts we drew in assignment 2. However, more female users mentioned the "Islamic Republic of Afghanistan" than other genders. We also noticed this occurrence in the second assignment and checked several posts containing the "Islamic Republic of Afghanistan," which related to "horrific attack," "policy," "economy," and "death." We suggested the analysis of this figure might lead to the further research question of whether females are more concerned, care about, and express their thoughts about human rights, freedom, and national peace in their private posts on social media.

Hovering the mouse in the location position in the plot will provide the total number of times that users mention that location. For example, "United States" was mentioned 501 times in the figure above. When clicking the proportion of the pie where "United States" is located, the figure below demonstrates further the distribution of genders of users mentioning "United States." There are 123 females, 307 males, and 71 others who wrote "United States" in their narratives.

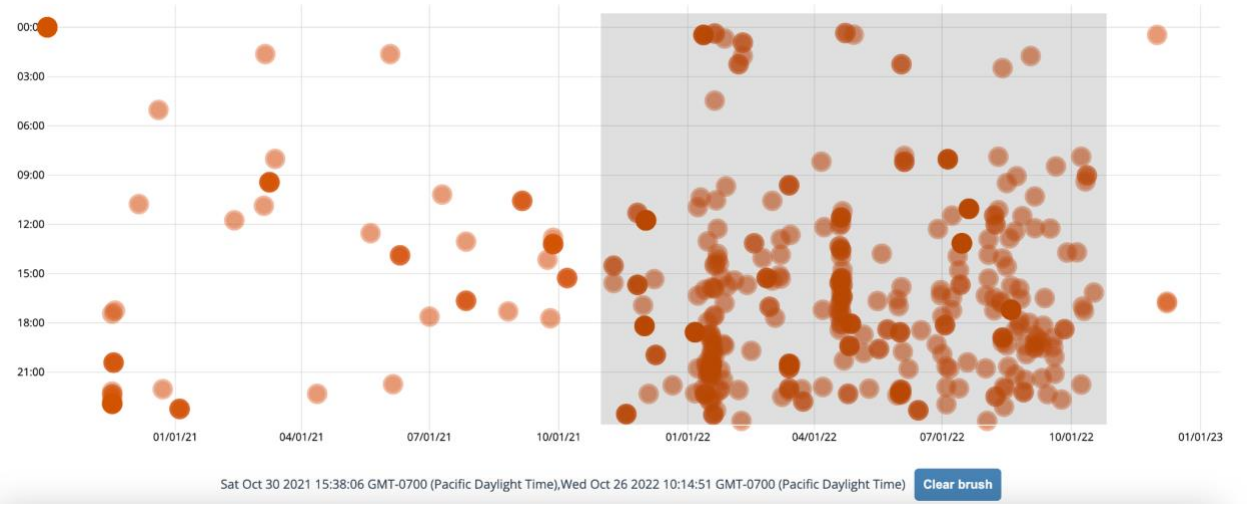


Timeseries

This visualization is selected as it answers the questions ‘at what time of the day do people tend to post toxic contents’ and ‘whether or not there appears to be more violations of Pixstory’s intention of building a clean environment in more recent days’.

We determined any post with a score higher than 0.5 from any of the following fields: Toxicity, Severe_Toxicity, Obscenity, Identity_Attack, Insult, Threat, Sexual_Explicit (all previously generated by Detoxify), should be a toxic post. The dataset is derived as one product of running ‘Extract Subsets.ipynb’ and saved as ‘toxic_posts.csv’ under the D3 folder.

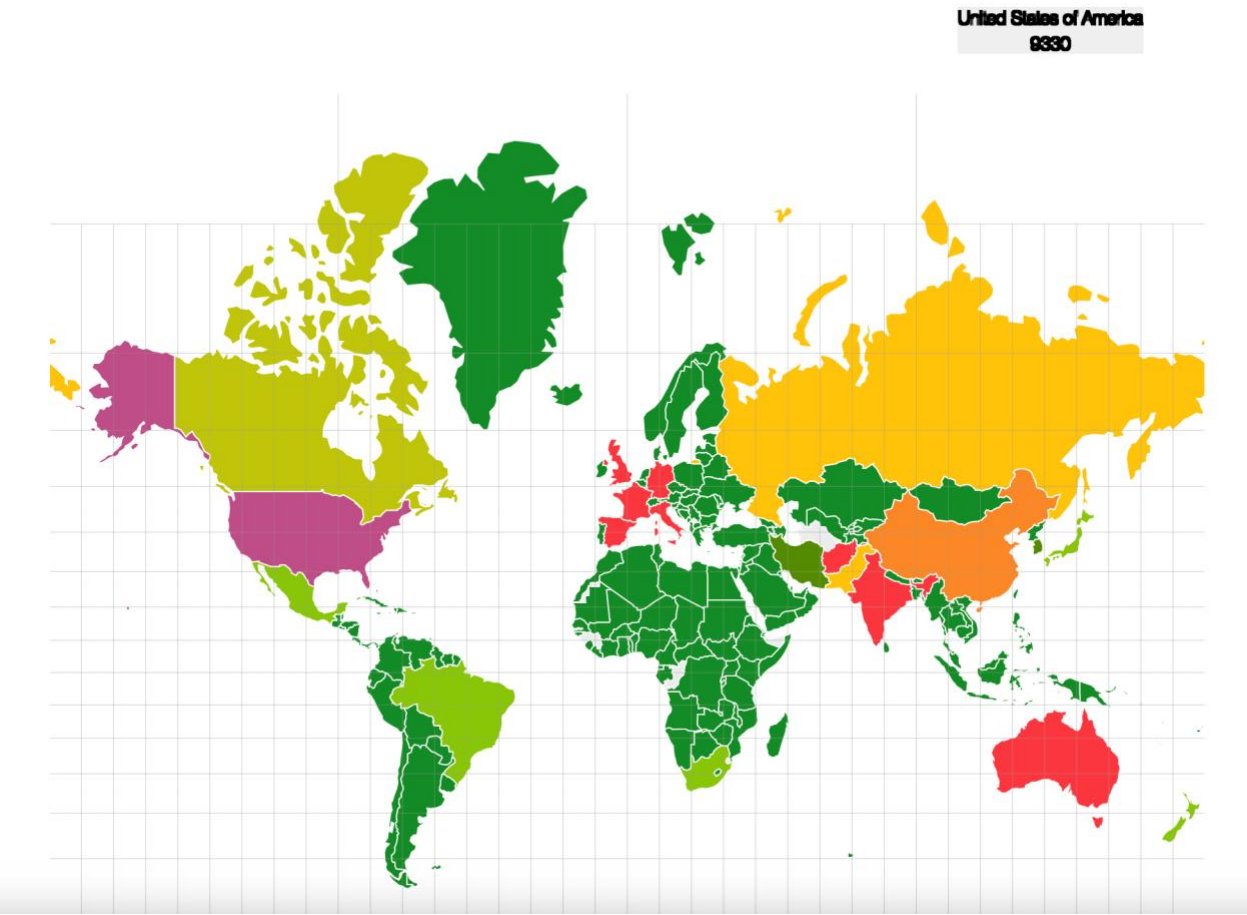
In total there are 518 toxic posts. Based on what we saw in the visualization, we claim that Pixstory users post more toxic contents during the period 12pm-3am. Particularly in between November 2021 and October 2022, there are numerous violations to Pixstory’s norms, which can induce further analysis on triggers of this phenomenon. However, since there are not many records starting November 2022, we fail to conclude that users tend to post more toxic things in more recent days.



Heatmap

We built 4 different heatmaps using occurrences of country names in posts to see the popularity of each country as well as how much hatred is targeting each one of them.

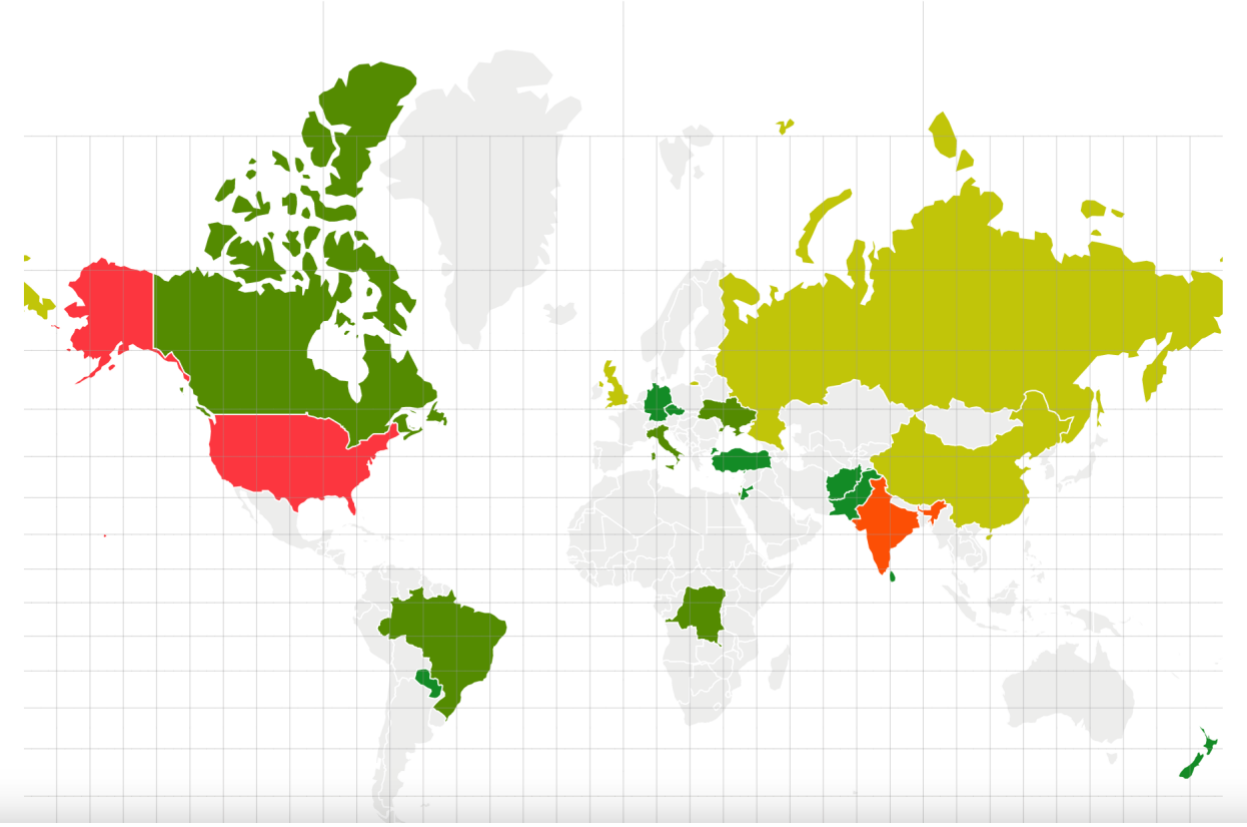
Our first heatmap shows how many times (both good and bad) a country is mentioned on the platform. As a result, the U.S. is the most frequently discussed country, and it was mentioned 9330 times. Others that attracted a lot of attention include Australia, India, Afghanistan, France, Spain, Italy, Germany, and U.K.. This aligns with socio-political situations around the world.



The rest 3 maps focus on negative attitudes targeting the countries. One is built using all toxic posts that have names in their contents, one uses records that are labeled as sarcasm, and the last processes posts that are determined to be hate speech.

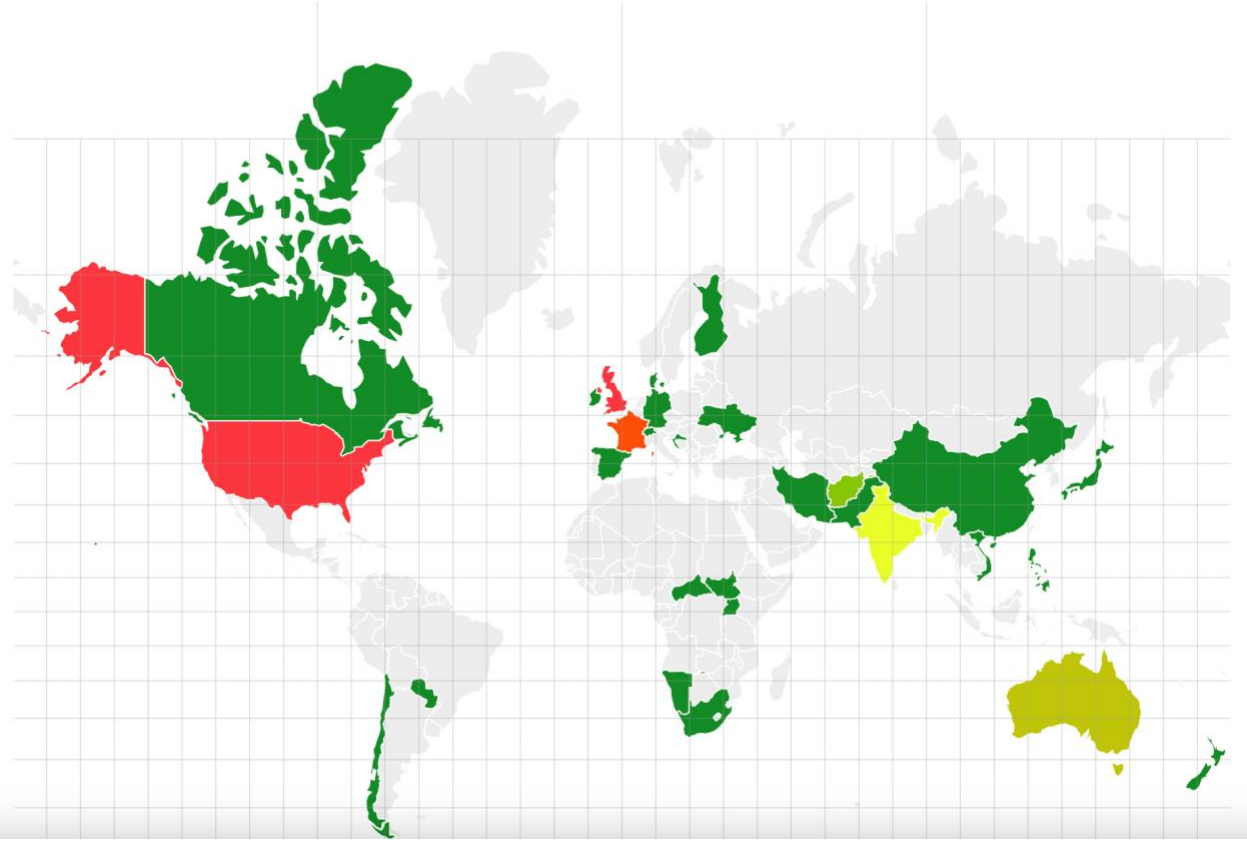
As for toxic posts, only two countries were mentioned at least 5 times, the U.S. (12 times) and India (5 times). All the occurrences add up to 58, which means around 11% toxic posts are directly targeting a country.

Hover over to see how many TOXIC posts have mentioned the country.

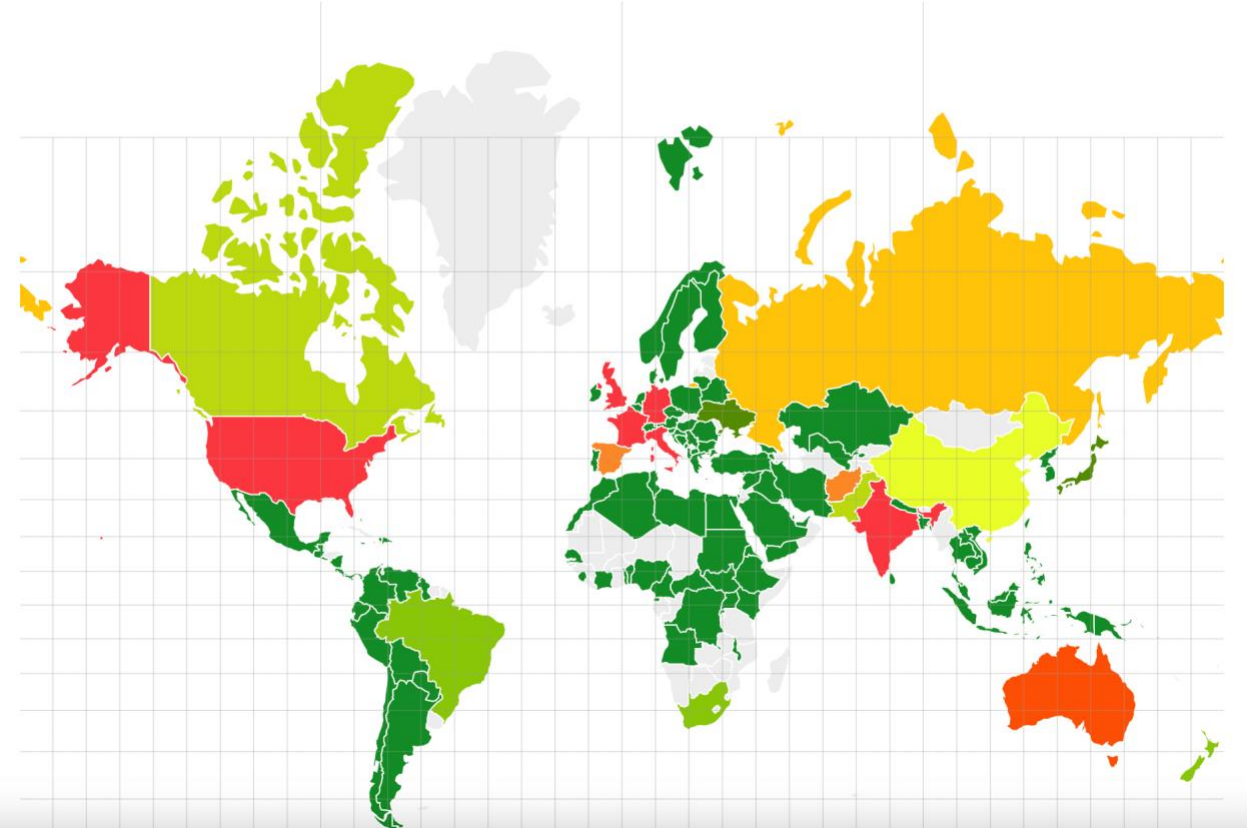


The sarcasm heatmap tells a similar story. Out of 134 mentions, the U.S. is accounted for 51 of them, and India is for 7 times. But this time, the U.K. (10 times) and France (9 times) also became frequently mentioned subjects.

Hover over to see how many SARCASTIC posts have mentioned the country.



What can be inferred from our hate speech map is that there are only two countries that are targeted more than 100 times, which are U.S. (355) and U.K. (219). One thing we would like to highlight is that hate speech is more of a concern for the Pixstory community, compared to toxic and sarcastic contents. We found 1840 hate speech posts from the whole dataset, which is a much larger volume compared to the other two.



In a nutshell, the countries that attracted more attention overall are more likely to be targeted/attacked by the users. Among 43101 mentions of country names, 2032 are problematic, which takes up around 4.71%.

2. Did Image Space allow you to find any similarity between the PixStory story images that previously was not easily discernible?

Image Space is a powerful tool for object detection and finding images with similar objects. It can identify a rhinocero on grass and match it with elephants in a grassy environment and in an amusement park. When we detect them with our eyes, we may miss the elephant mixed in the colorful background.

Image Space is also proficient at categorizing and drawing conclusions from images. For instance, when given a photo of a person, it can output images with people of different genders, ages, and races. Similarly. When given an image of a plate of salad, it can return images of rice,

french fries, and oatmeal. This tool is often more efficient than humans in categorizing similar objects.

Moreover, it can establish connections in a broad sense. When presented with a photo depicting a countryside scene, it can show views of the beach, buildings, and walls. Unlike humans who might miss images that show landscapes not closely related to the countryside, such as a wall with sunlight passing through, Image Space is able to identify them.

3. What type of location data showed up in your data? Any correlations not previously seen, e.g., from assignment 1?

Based on the images generated from Geoparser, we conclude that data points are primarily concentrated in North America, Europe, and Southeast Asia worldwide. When focusing on the United States, we observe that data points are denser closer to the ocean than in inland areas. Specifically, data points are concentrated in big cities such as Los Angeles and Washington. This trend is also visible when we zoom in on California, where we see a higher density of data points along the coast, particularly in cities such as Los Angeles, Long Beach, Irvine, and Oakland.

These findings align with the conclusions we drew in Assignment 1, which showed that the top three percent of languages used in the posts were English (70.9%), Italian (12.3%), and Bengali (2.9%). Given that North America and Europe are areas with higher density data points, it follows that English and Italian languages are predominant in Pixstory narratives. It is a direct and clear visualization of Geolocation.

Combining mentioned country names with toxicity, hate speech, and sarcasm labels, we got a set of countries: U.S., U.K, India, and more, that are more frequently targeted/attacked by users, as discussed in section 1.

Also include your thoughts about Image Space and ImageCat – what was easy about using them? What wasn't?

Image Space and ImageCat are software tools that can be run from dockers and have user-friendly interfaces, making it easy to install, set up, and analyze images. By entering images, we can search for similar images. Image Space web page displays all similar images for us to visualize and compare.

To improve efficiency, it would be useful if these tools not only output similar images but also sort them by similarity. For instance, when searching for a mosaic, the tool returns images of other mosaics and charts. While bar charts have some similarity to the original image, they are

not the closest match. Sorting the images by similarity would enable faster comparison and analysis, for example when looking for images within a certain similarity range.

At the same time, we have encountered an issue with the container deploy-imagespace-mongo-1, which does not run properly on Windows systems. The container exited and had a black log. To be more accessible to a broader range of systems, including Windows, Image Space and ImageCat would benefit from improved compatibility.