

DSCI 550

03/7/2023

Professor Mattmann

Assignment 1

Team 12: Jai Agrawal, Daniil Abbruzzese, Todd Gavin, Tania Dawood

## **Pixstory Insights**

Our group worked on exploring the impact Covid had on social media usage. In addition to the Pixstory dataset, we chose to use a Snapchat Average Daily Users dataset, a COVID-19 data and a YouTube trending videos dataset. Each of these datasets had different MIME Types as stated in the assignment 1 document and by merging all the datasets into one master dataset, we were able to answer the research questions which examined trends and patterns within our data.

### **Our Process:**

As the first step for this assignment, we added 5 features to the Pixstory dataset. These included "Interest," "Sports Events," "Associated Film Festivals," a "Hate Speech Flag," and "Hate Speech Detected" as well as a Sarcasm flag. This was done by creating Python functions that use BeautifulSoup to scrape websites like GLAAD to detect hate speech, ADL.org to find hate symbols, Sporting Events from 2020-2022 which were found from [www.topendsports.com/events/calendar](http://www.topendsports.com/events/calendar), Film Festivals from 2020-2022 which was obtained from [www.filmfestivaldatabase.com](http://www.filmfestivaldatabase.com) and a sarcasm corpus from [nlds.soe.ucsc.edu/sarcasm1](http://nlds.soe.ucsc.edu/sarcasm1). All the scraped data from each website was saved to its own dataframe before being merged into a master dataset. Our hate speech and sarcasm flags were added as boolean columns and if either hate speech or sarcasm was detected in any of the posts, the respective column denoted 1 and if there was no hate speech or sarcasm was detected, the column denoted a 0. This was done by obtaining a list of sarcastic phrases and hate speech words and comparing it to the Pixstory dataset. For this, we used an edit distance algorithm and flagged a post if the edit distance score was over 50%. Lastly, by analyzing these datasets together we were able to pull a number of insights. We did this by creating Python scripts to merge the datasets as well as by using the packages that were referenced in the assignment 1 task file. By doing this, we were able to identify trends related to our goal.

The next step of our process was to identify three additional datasets and merge them into the master dataset which already had all the Pixstory data and our features. However, in order to find additional insights from our selected datasets, we came up with three additional features to each.

### **Datasets we chose:**

Dataset #1: Snapchat Daily Average Users

In order to create this dataset, we obtained Snapchat's quarterly revenue and daily average users from Statista and its daily stock value from Yahoo Finance. Combining the three datasets, we were able to create our own dataset with a MIME type and subtype of Text/ CSV. Our data

ranges from 2020-2022 as that is the range of data we have available for the Pixstory dataset. For the Snapchat dataset, the features we will be using for analysis are:

- Daily Average Users
- Stock Price
  - The stock data has a few missing values, since it only has the values that occur on the days during which the Stock Market is active. We'll fill these in with the value of the last closing price.
- Revenue

#### Dataset #2: COVID API

As we are looking at data during the years 2020 - 2022, our group thought it would be necessary to look into COVID data to check if that had any impact on how social media was being used. The MIME Type of our COVID dataset is Application/ JSON and the features of this dataset that we took into consideration were:

- Daily cases
  - This feature was derived from taking data on cumulative COVID cases in India on a daily basis (source: (<https://covid19api.com>) and finding the difference from the prior day
- Daily vaccinations
  - This feature was derived from taking data on the cumulative vaccinations in India on a daily basis (source: <https://github.com/owid/covid-19-data>) and finding the difference from the prior day
- Daily Deaths
  - This feature was derived from taking data on cumulative deaths in India on a daily basis (source: (<https://covid19api.com>) and finding the difference from the prior day

However, this data set only had data available as early as 1/15/2021, which listed India as having 0 vaccinations against COVID. Because the Pixstory data set starts 1/12/2020, we decided to add this missing dates and input 0's for all vaccinations. This was justified because according to this data set India hadn't had any vaccinations until 1/16/2021.

#### Dataset #3: Top YouTube Daily Trending Videos

This is a dataset of the top trending videos on YouTube on any particular day. The MIME Type of this dataset is Video/ MP4. The data ranges from 2020 - 2022 and the features of this include:


- Highest trending video name,
- Highest trending video channel,
- Highest trending video category,
- Highest trending video views,
- Highest trending video likes

Once we were able to merge all our datasets together with the appropriate features we mentioned earlier, we were able to use Apache Tika to run a similarity analysis and derive certain key insights.

## Key insights:

To start our analysis, we divided the 95,000 JSON files into different folders, each containing 100 files each. We were then able to run a Tika Similarity analysis using distance/ similarity metrics like Jaccard Similarity, Edit Distance and Cosine Similarity. However, we found that the clusters that made the most sense to us were those created by Jaccard Similarity and Edit Distance. The Jaccard Similarity allowed us to create clusters based on age and gender, whereas the edit similarity gave us clusters based on the different monthly revenues found. However, the cosine similarity only showed us two clusters which we did not think were helpful in our analysis. (Question 7)

categorical colors range [1: 20]



1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

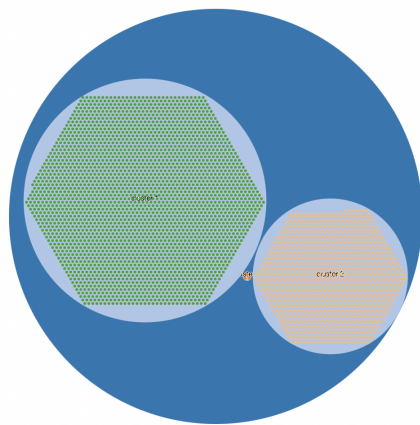



Figure 1 - Cosine

categorical colors range [1: 20]



1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

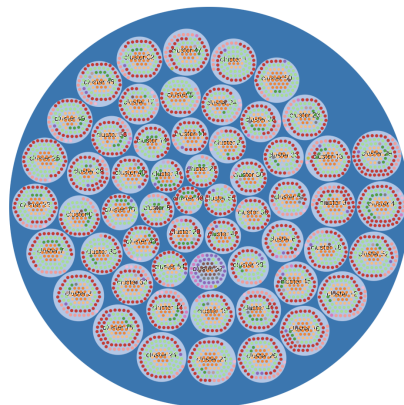
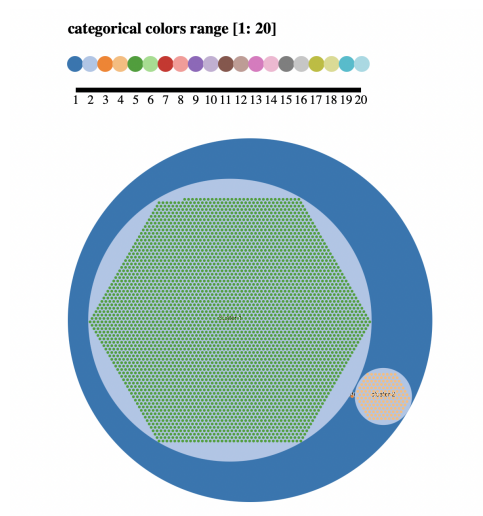
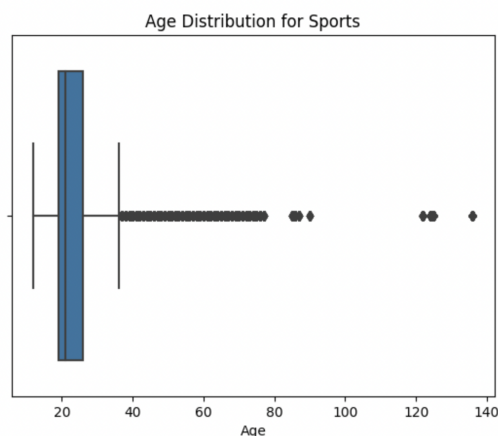


Figure 2 - Edit

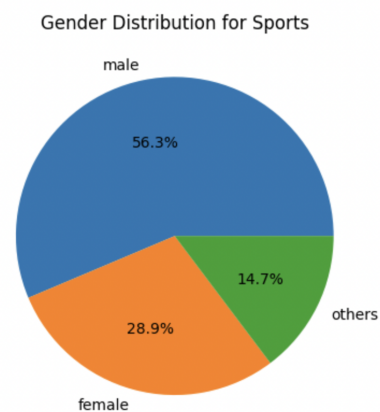


**Figure 3 - Jaccard**

However, once we had our data sorted, we were able to figure out the most popular topics that were discussed in Pixstory posts. These included: sports, health, entertainment, food and politics. After running a Tika Similarity analysis, we found the average age for users posting content related to the most trending topics was between 18-30 years old, which aligns with the general social media user age. However, further analysis showed that there was a slight gender disparity when it came to posts related to top interests. We noticed that there were generally more males posting about those topics compared to females. Below we have shown our visualizations for the sports topic for reference (Question 1 and 3)

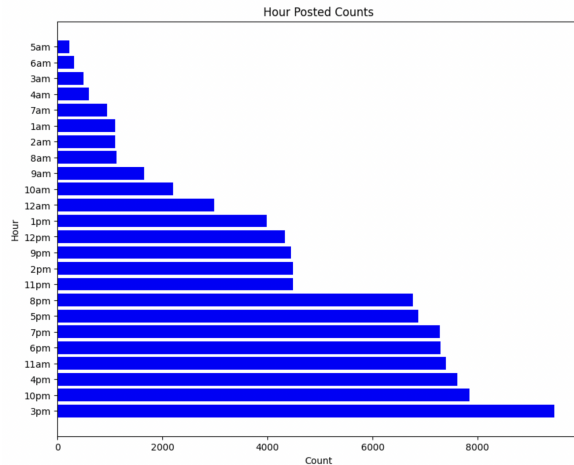


**Figure 4**



**Figure 5**

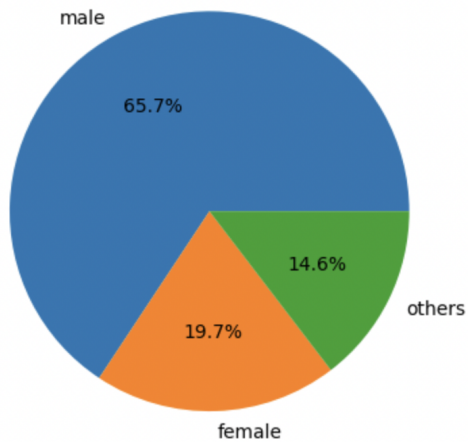
As for whether the time of day of the post matters, we found that the afternoon to evening is when most activity took place on Pixstory. Our analysis shows a majority of posting happens between the hours of 11 am to 8 pm. However, an interesting find was that posting during “lunch hours” 12 pm to 2pm was significantly lower. (Question 2)



**Figure 6**

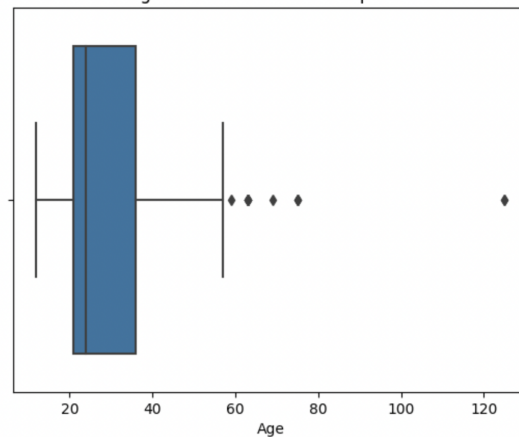
Another interesting insight we were able to find by analyzing Pixstory data related to hate speech was that people within the age range of 21-37 were more likely to post content that was flagged for hate speech. Our group particularly found this insight interesting as the range we found falls within the Millennial and Gen X generations, and we know that there is a current social awareness trend that Gen Z has particularly become akin too. However, it is also important to note that the majority of posts that were flagged as hate speech came from males. Our data shows that 65.7% of the posts flagged as hate speech were created by males and only 19.7% were created by females. (Question 4)

**Gender Distribution for Hate Speech**



**Figure 7**

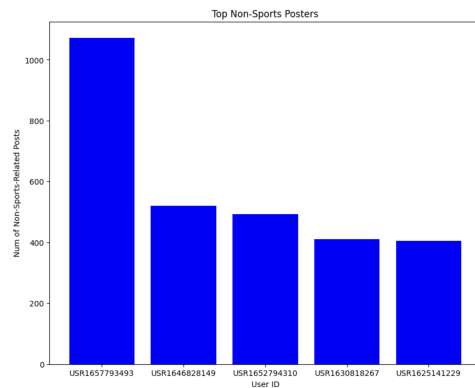
**Age Distribution for Hate Speech**



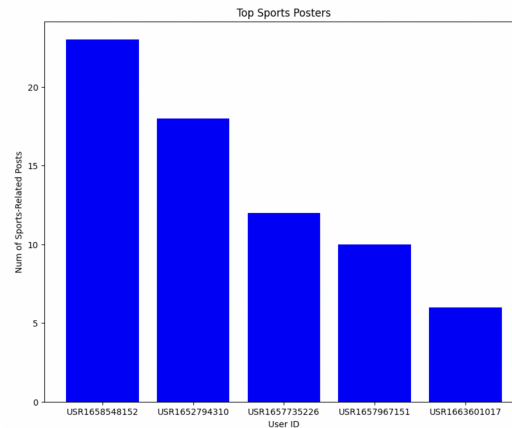
**Figure 8**

As for frequently co-occurring features that defined a particular set of users, we discovered that there were two types of users posting to Pixstory on the days there are major sporting events. The first type being users who enjoy posting and engaging in sports related content and the

second type are users who post and engage with non-sports related content on days where there are many sporting events. We decided to classify these types of users as sports enthusiasts and non-enthusiasts. We were able to get these two types of users by creating two additional columns for our dataset. These were “Number of Sports Events” and “Top Sports Posters.” By conducting this analysis, we were able to find the top 5 sports enthusiasts and non-enthusiasts as shown below. (Question 5)



**Figure 9**



**Figure 10**

Lastly, when looking at the “indirect” features we extracted from the data we made the following discoveries:

For the Snapchat dataset, when we compared Snapchat daily average users (DAU’s) to Pixstory DAU’s, we found that there is not much correlation in the trends of the two or its revenue collections. While this is not a solid answer to our question, we can interpret this data to say that users generally being online on other social media platforms (such as Snapchat) do not necessarily translate into users being online on PixStory. An interesting find here is that Snapchat’s DAU trends the same as its revenue. Additionally, we found that PixStory’s daily post count does not correlate with Snapchat’s stock value. This means that the market’s perception of social media (such as Snapchat) may not imply increased user engagement on other social media (such as PixStory). Further, we can then say that as the stock price for other social media (Facebook, Snapchat, Twitter) rises, it may not mean much for PixStory.

As for the YouTube dataset, we found that the days on which there were more likes and views on trending videos, there seemed to be more activity on Pixstory. However, we cannot comment on the causation here as our data was limited. Rather, we believe that there is some correlation between the potential content of the video, and activity that is going on in the real world. For example, a YouTube video could be related to a sporting event currently taking place, and there could be an increase in Pixstory posts about said sporting event.

And for the COVID API dataset, we found that the new daily Covid cases, led to more covid related posts as well as a rise in “India” related posts as our dataset is mainly related to activity

in India. We also believe that as Pixstory is an Indian app, it is highly marketed in that region and more likely to have posts related to happenings in and around India. (Question 6)

### **Unintended Consequences:**

As we learnt in class, posting on certain social media platforms (Snapchat, YouTube etc) can lead to unintended consequences such as a decline in mental well-being due to the absence of measures that ensure a safe and less toxic environment. Negative comments can surface on platforms like YouTube, without algorithms in place to prevent them from appearing under videos and YouTube shorts. Constant exposure to negative feedback, particularly when multiple comments are read, can have an adverse impact on both content creators and their subscribers. Our COVID and snapchat datasets reveal that during the peak of the pandemic, social media usage was at an all-time high as people were confined indoors with limited social interaction and fewer entertainment options.

Additionally, another unintended consequence of these datasets could be digital targeting. The Pixstory dataset includes features such as age and gender which could lead to targeted ads, preventing users from seeing similar products from different manufacturers and for cheaper prices.

### **Conclusion:**

As first time users of Apache Tika, we initially faced some difficulties in setting it up as well as conducting a similarity analysis. However, once we were able to figure out how to correctly install and use Tika, our group found that Tika was a great resource for pulling insights from large datasets with ease.

Overall, we believe we were able to uncover some interesting insights on social media usage and the increase or decrease of COVID-19 numbers. For example, every time there was an increase in the number of daily COVID cases, the amount of posts to Pixstory increased significantly.