Report – Assignment 1 Team 3

Jimin Ding 1556886350
Xiaoyu Dong 2468117466
Hui Qi 3206742781
Mingyu Zong 3484496941

Abstract

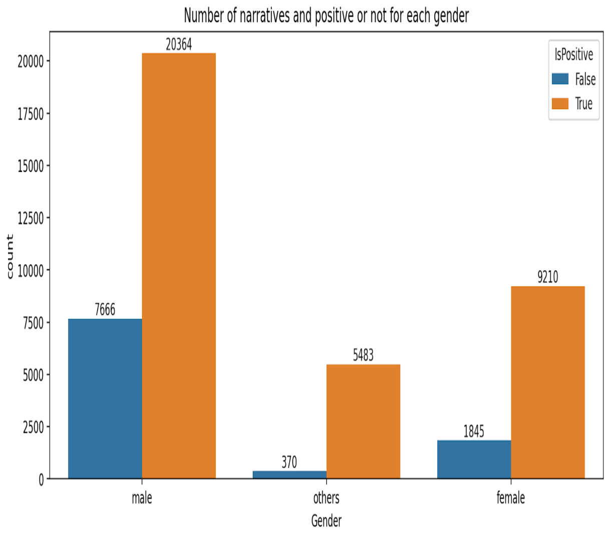
This report investigates features from big data of the PixStory application from January 2020 to December 2022, combined variables from different MIME types of files, and the overall data similarity to analyze users' behavior information and classify their patterns. The final dataset has over 95,000 instances and 28 variables. Through analyzing the big data, we want to provide thoughtful suggestions on creating a better environment for users to post positive posts in this PixStory application and contribute to stricter detection for certain groups of users with a higher probability of posting hate or sarcasm posts.

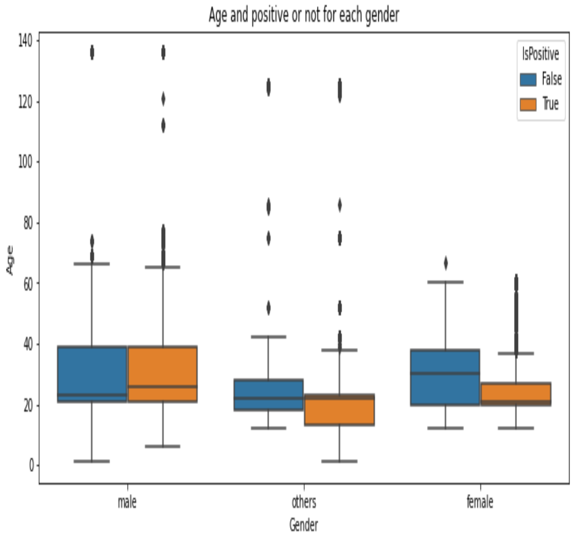
Features from new joined datasets

We chose three non-text datasets to add new features on positive word identification, public holidays, and global covid-19 data.

Dataset 1:
For this dataset, we selected three positive words - "agree", "good", and "fun" - and combined them into three images of MIME type, image/png. We used the Tesseract tool to extract text from images and compared it with the PixStory narratives. We changed both positive words and narrative to lower cases for a direct match. If a positive word existed in a narrative, we identified it as containing positive words and gave it a True value for Positive Words.

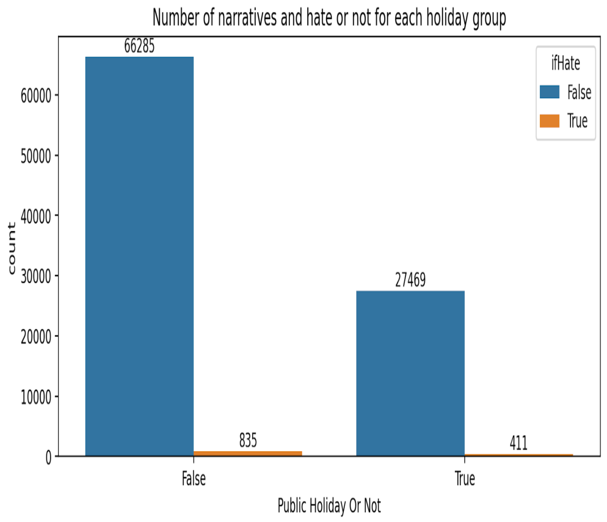
We also assumed that gender and age are related to the number of positive posts. We found that female PixStory users had a higher tendency to make positive posts. Additionally, the effect of age on positive posts varied by gender. For male posters, the median age for those making positive posts was larger than those making neutral or negative posts. For female posters, the median age for those making positive posts was lower than the other group.

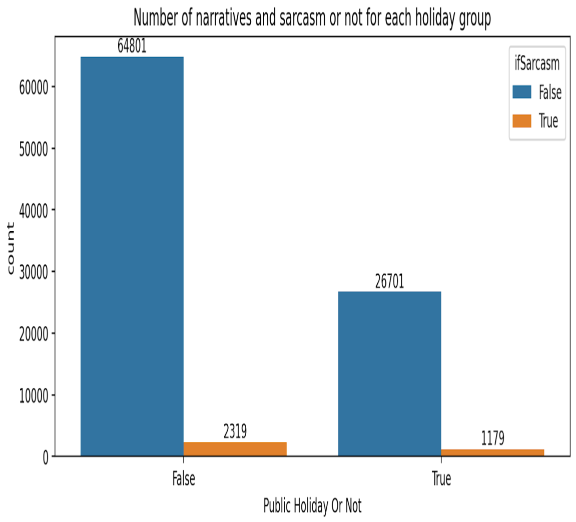




Dataset 2:
In this dataset, we used a public holiday API to find the date the user made the post a holiday in each continent or not. Its MIME type is application/API+JSON. We sent a URL query to the public holiday API and received a JSON response with keys: “date", "localName", "name", "countryCode", "fixed", and "global". The last two keys were boolean values indicating whether the holiday was fixed or global. We used the request library to parse the API response and determined whether the date users made posts was global public holidays or not.

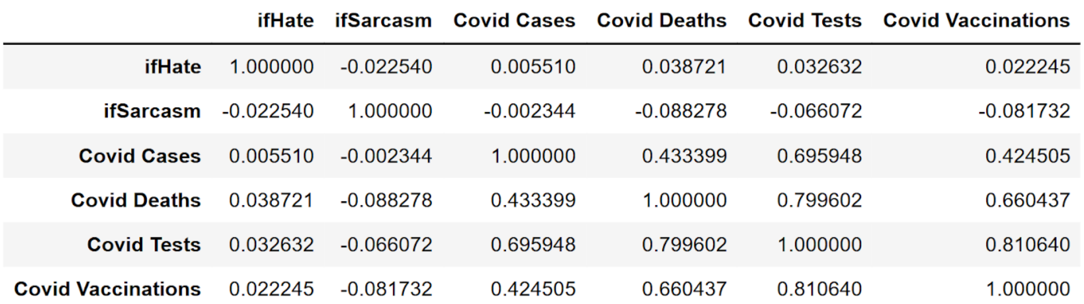
Our initial assumption was that users would have a lower rate for hate and sarcasm posts on public holidays when they are more relaxed, but our results showed the opposite. Users had a higher rate for hate and sarcasm posts on public holidays, possibly because they were more excited and emotional.





Dataset 3:
For this dataset, we collected numbers of cases, deaths, tests, and vaccinations in a day globally. The MIME type was application/zip. We extracted the zip file, unzipped it and got a CSV file for Covid data. We used the Pandas library in Jupyter Notebook to parse the data, sum each country's data and get a daily global overview. We then compared cases, deaths, tests, and vaccinations with whether the post was classified as "hate" or "sarcasm" or not.

Our prediction was that when there are more daily increases in Covid cases or deaths and users are worried about Covid, there will be more hate and sarcasm posts. However, our results showed a positive but very weak correlation between Covid data and hate or sarcasm posts. This may be due to the fact that Pixstory users have a negative attitude when Covid is more serious overall, with more deaths and vaccinations. However, the correlation was too weak to be statistically significant.

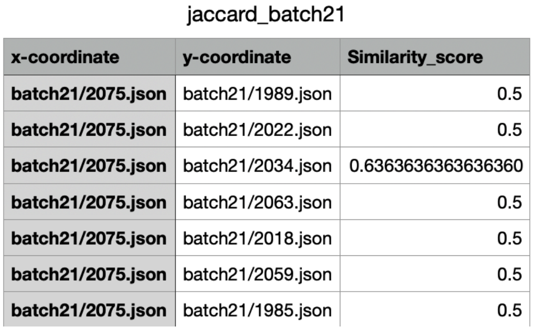


	ifHate	ifSarcasm	Covid Cases	Covid Deaths	Covid Tests	Covid Vaccinations
ifHate	1.000000	-0.022540	0.005510	0.038721	0.032632	0.022245
ifSarcasm	-0.022540	1.000000	-0.002344	-0.088278	-0.066072	-0.081732
Covid Cases	0.005510	-0.002344	1.000000	0.433399	0.695948	0.424505
Covid Deaths	0.038721	-0.088278	0.433399	1.000000	0.799602	0.660437
Covid Tests	0.032632	-0.066072	0.695948	0.799602	1.000000	0.810640
Covid Vaccinations	0.022245	-0.081732	0.424505	0.660437	0.810640	1.000000

Additional datasets' "unintended consequences" and goals of the platform

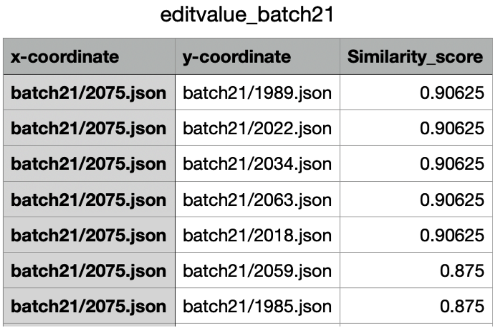
PixStory is a relatively less toxic social media platform, with a high proportion of positive posts and a low incidence of hate speech. By analyzing datasets, we can identify relationships between user demographics, holidays, COVID, and attitudes towards PixStory posts. We can pay special attention to particular groups of users or times of the year to create a more welcoming environment. For instance, younger male users and those posting during holidays may have a greater tendency to post negative content, but Pixstory can detect these patterns and take action to promote positive online behavior.

Tika Similarity



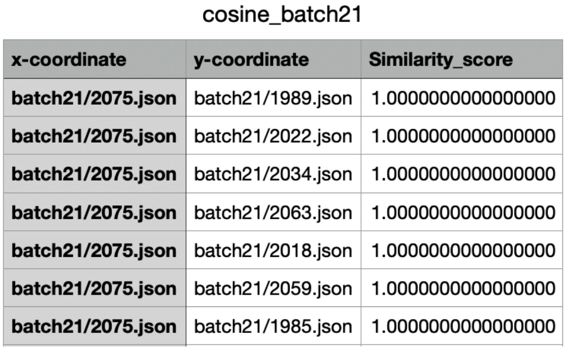
jaccard_batch21

x-coordinate	y-coordinate	Similarity_score
batch21/2075.json	batch21/1989.json	0.5
batch21/2075.json	batch21/2022.json	0.5
batch21/2075.json	batch21/2034.json	0.6363636363636360
batch21/2075.json	batch21/2063.json	0.5
batch21/2075.json	batch21/2018.json	0.5
batch21/2075.json	batch21/2059.json	0.5
batch21/2075.json	batch21/1985.json	0.5



editvalue_batch21

x-coordinate	y-coordinate	Similarity_score
batch21/2075.json	batch21/1989.json	0.90625
batch21/2075.json	batch21/2022.json	0.90625
batch21/2075.json	batch21/2034.json	0.90625
batch21/2075.json	batch21/2063.json	0.90625
batch21/2075.json	batch21/2018.json	0.90625
batch21/2075.json	batch21/2059.json	0.875
batch21/2075.json	batch21/1985.json	0.875



cosine_batch21

x-coordinate	y-coordinate	Similarity_score
batch21/2075.json	batch21/1989.json	1.0000000000000000
batch21/2075.json	batch21/2022.json	1.0000000000000000
batch21/2075.json	batch21/2034.json	1.0000000000000000
batch21/2075.json	batch21/2063.json	1.0000000000000000
batch21/2075.json	batch21/2018.json	1.0000000000000000
batch21/2075.json	batch21/2059.json	1.0000000000000000
batch21/2075.json	batch21/1985.json	1.0000000000000000

Jaccard Similarity computes the portion of common words in two documents in the union of unique words from each file. Regarding this project's JSON files, it considers counts of similar features instead of words. Most of the files from our three test batches are assigned a score of 0.5 or 0.636, and a few get a score of 0.8. Given that our full dataset has 28 distinct features, a score of 0.5 indicates 14 similar features in the two input files. A high enough score will help us confidently find duplicates or near duplicates. However, in this case, it fails to offer enough information to distinguish users from generated groups. Summary statistics concerning features and clusters can be found in Step7/Jaccard/Jaccard_batch*.ipynb.

Edit Distance Similarity assigns all test files a score between 0.75 to 1. By definition, it takes both file length and word position into consideration. Because of the diverse post contents and the varying COVID-related numbers, it is reasonable that any pair of input files would be marked as dissimilar. Among all the features, the 'Narrative' one contributes to the most significant portion of dissimilarity, as its length varies a lot and can contain all kinds of vocabulary. For this project, since it considers an input file as a whole text instead of a collection of features, it also does not provide enough feature-wise insights for users in different groups.

Cosine Similarity pays attention to the words' occurrence and the number of occurrences. Theoretically, this is the best measurement for finding duplicate documents or determining how similar two files are. Our group members attempted the identical three batches. However, most of the scores we obtained were 1, even though the input JSON files differed. The rest were extremely close to 1. Since the dataset has discrete and continuous features, we hypothesize that the script is somehow comparing only metadata or has trouble processing discrete features, which further causes its unexpected output.

Thoughts about Apache Tika

Apache Tika is a powerful parser with fast processing speed and a wide coverage of document types. With its functionality, when handling rare file types, one may not need to install corresponding applications to get data out of the file. Also, evidence shows that processing 95000 entries would typically cost 10 to 15 minutes. This feature allows it to be a competitive parsing tool in the era of Big Data. On the other hand, installation and configuration of this tool is fairly not user-friendly, especially when someone is not directly using the GUI or the server jar. It is easy to run into connection issues.

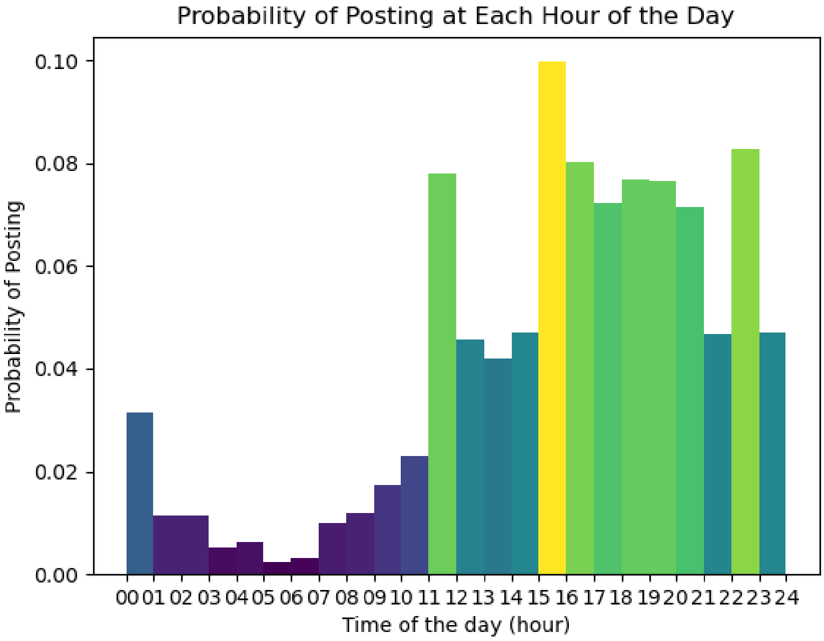
Deep thinking

1. Are there clusters of users with similar features that tend to post about the same topics?

According to the statistics of Jaccard clusters in corresponding notebooks, Jaccard clusters created from one test batch tend to share the most popular topic. For batch 21 and 181, it is "Technology, History, Food, Entertainment, Sports, Environment, Science, Inequality, Education, Health, Politics, Economy, Climate change", and for batch 565 it is "star wars, obi wan kenobi". Similarly, most popular topics for posts within the same batch are shared by clusters with Edit Distance similarity. If we could run the script on the entire pixstory dataset, one of the measurements may have the potential to group all posts by topics they focus on.

2. Does the time of day of the post matter?

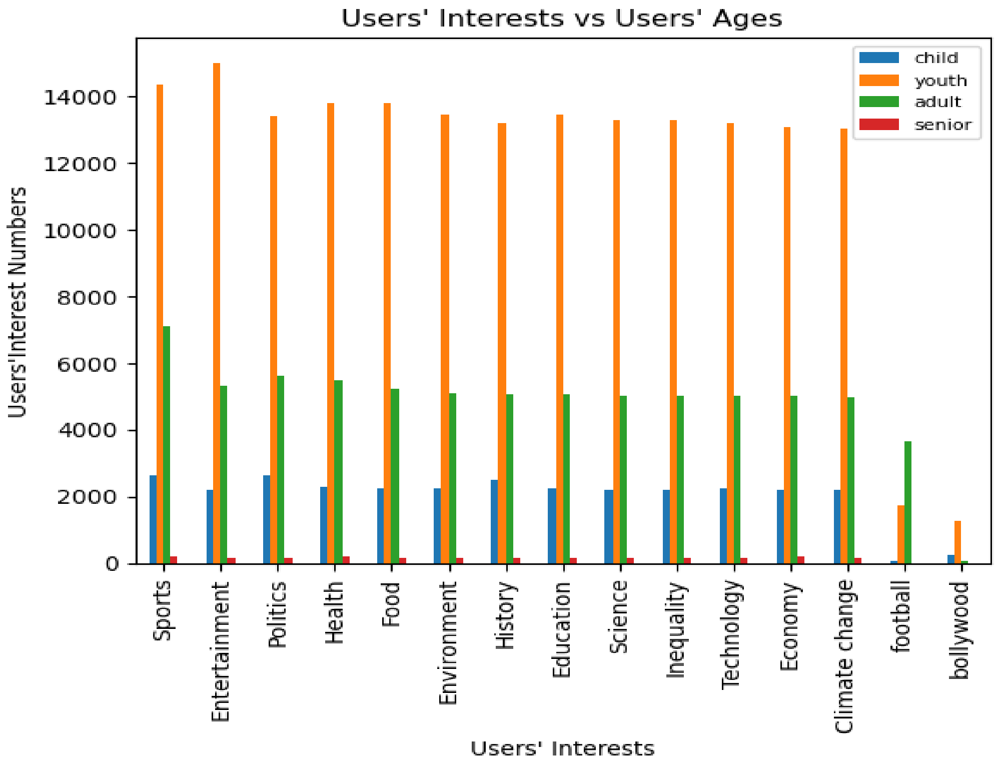
We drew a histogram illustrating when users sent the posts in the PixStory application. According to the plot below, users are more likely to post from 11a.m to 12 p.m. and from 3 p.m. to 11 p.m. on a random day. Users are less likely to post from 1 a.m. to 10 a.m. In the figure, different colors represent different probability levels of sending a post to help us detect the pattern.



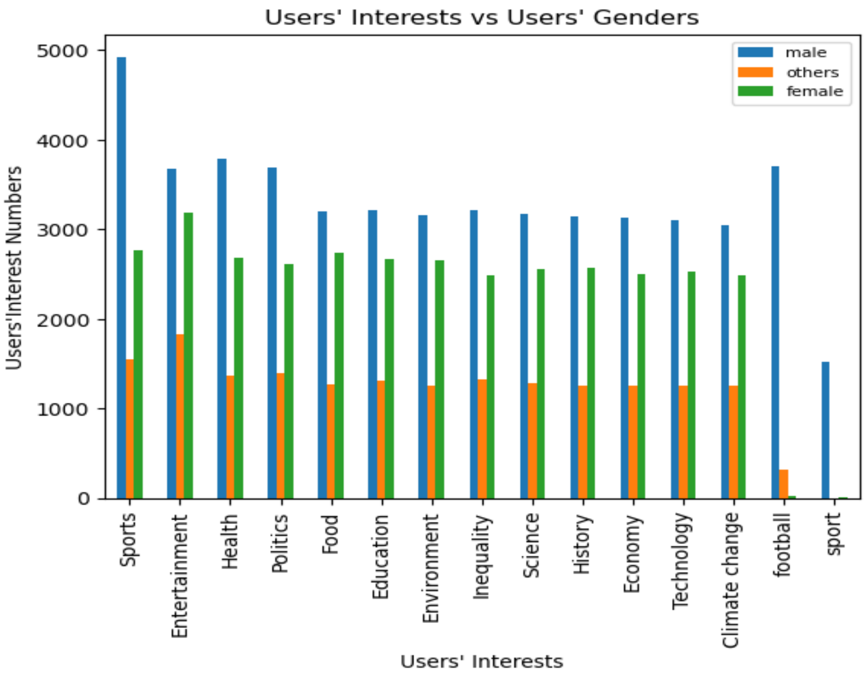
3. Are specific ages or genders of the users more likely to post about specific topics?

We are interested in the relationship among users' interests, genders, and ages.
We classify users into four age groups: ages between 0 and 14 are children; ages between 15 and 24 are youths; ages between 25 and 64 are adults; ages between 65 and 100 are seniors. Some users with ages over 100 are removed from this analysis since this data is spurious, as some users might set false and exaggerated information. Among all users' interests, I pick the 15 interests that most users have.

From the figure below, most users of the PixStory application are youths and adults. Youths have a wide range of interests because the counts of interest topics are considerably large and are close to 14000, except Bollywood. At the same time, youths' favorite topics are sports and entertainment. We notice that football is also marked as an "interest" tag while having the existence of the "sports" tag. It might indicate that football is the most topical and attractive sport or that some significant football games happened during the timeline of this dataset. Adults are most interested in sports and politics and also have broad interests. Moreover, this bar chart implies that children pay the most attention to sports and politics, which is the same as adults' notice and hobbies. It might indicate that children nowadays are concerning political things happening around them and relating to them. It could also be because some users enter false ages in their account information. Identifying which users input fake data is much more complicated than removing users' information with age periods over 100 from this analysis. So, we keep the data though we know there might be incorrect information. Furthermore, the PixStory application has fewer senior users than other age groups.



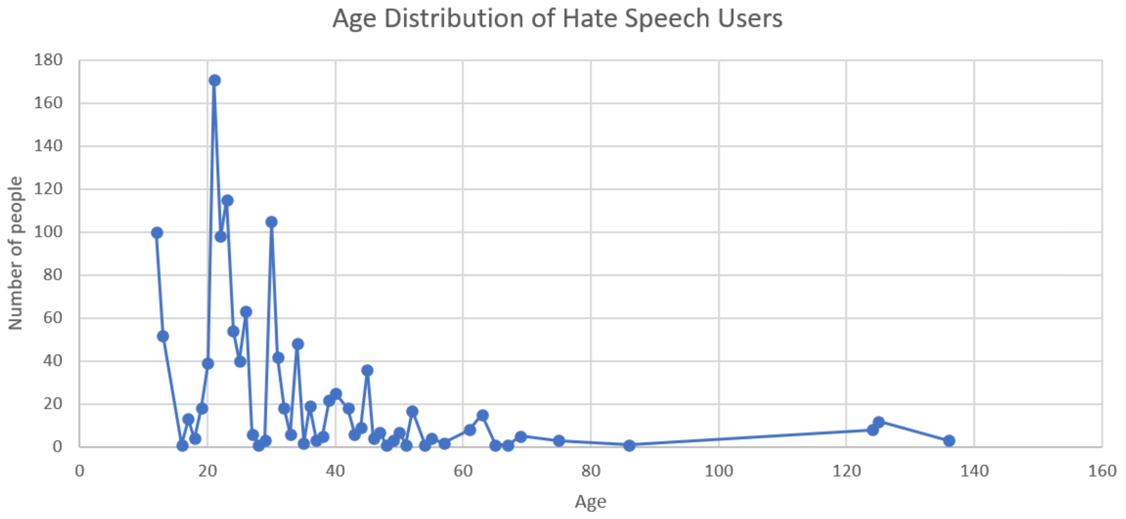
The figure below indicates that males are the largest group of users' genders. Males are primarily interested in sports, especially football. Males contribute a lot to the topic of football. The interest number in football for males is close to 4000, while it is less than 500 for females. Females are interested in entertainment and sports. It might imply that females favor sports other than football.



4. Are specific ages or genders of the users more likely to post any hate speech? Did you detect any?

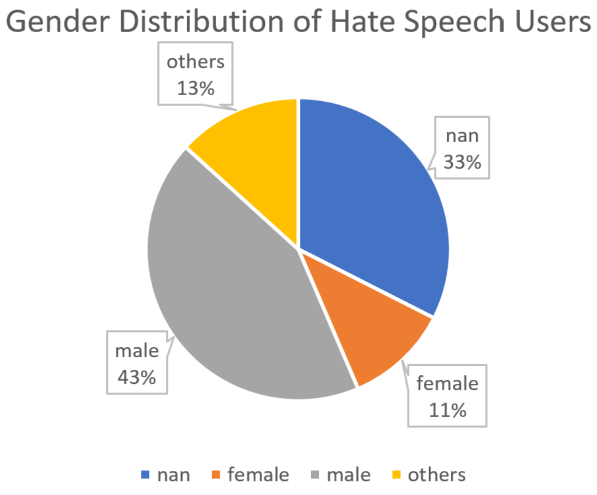
We collected the age and gender information of all the users who had posted hate speech posts (1,246 total), and plotted the corresponding statistics.

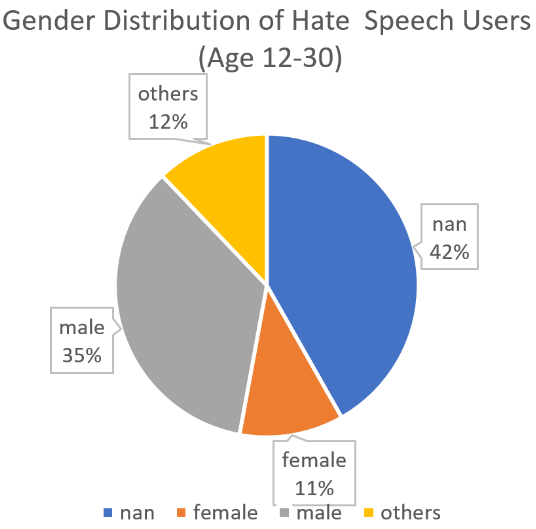
First, about the relationship between user age and hate speech , we found that 70.87% of the posts that contain hate speech were posted by users aged between 12-30. Young adults (18-25) made up the largest proportion of hate speech, at 43.26%.



As for the relationship between hate speech and gender, we found that among the users who posted 1,246 posts containing hate speech: male users account for the largest proportion, as 43.18% with 538 posts in total. Female users account for 11.08% and posted 138 articles; the proportion of other genders was 13.24%; and the remaining users did not add gender, accounting for 32.51%.

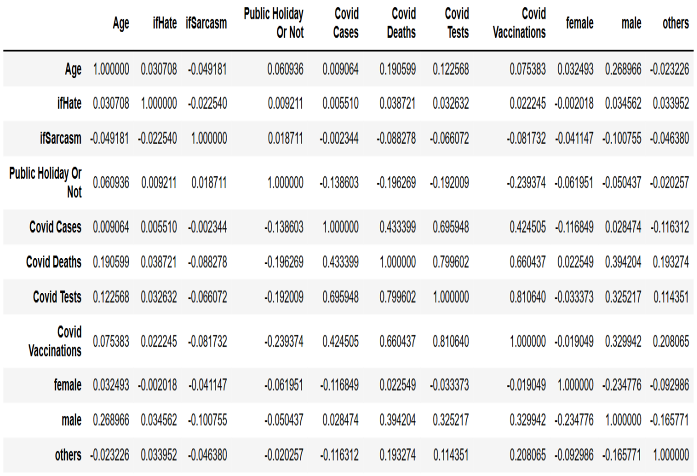
In addition, male users still make up the majority of those aged 12 to 30 who are the most likely to post hate speech, accounting for 35.00%. Women and other genders still account for a smaller proportion of users, at 11.10 % and 12.12 %.



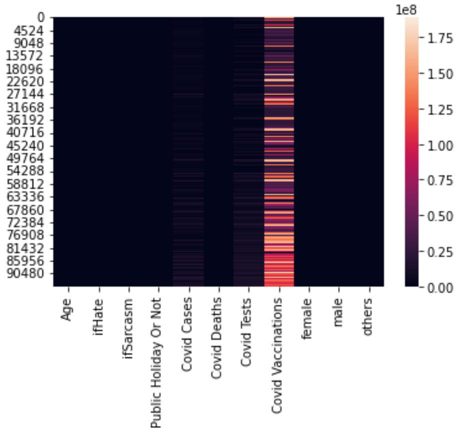


5. Is there a set of frequently co-occurring features that define a particular user or class of user?

We drew a correlation table and heatmap to find relationships between features. The relationship was weak overall in the PixStory dataset. There was a weak relationship between age and Covid death, with a correlation coefficient of 0.19. The correlation coefficient between age and male was 0.27, indicating that male users in PixStory were older than female users. Elder users have a higher covid death rate and higher possibility to be male. The correlation coefficient between public holiday and Covid vaccination was 0.27, indicating that there were more vaccinations on holidays. The relationship between hate or sarcasm and gender was very weak. We did not find evidence that male or female users had a higher tendency for hate speech.

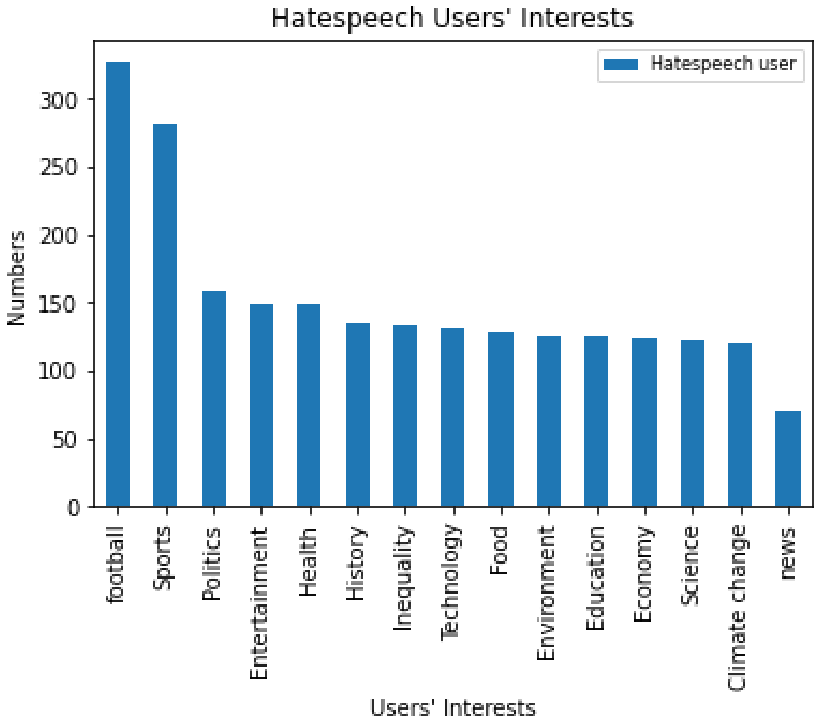


	Age	ifHate	ifSarcasm	Public Holiday Or Not	Covid Cases	Covid Deaths	Covid Tests	Covid Vaccinations	female	male	others
Age	1.000000	0.030708	-0.049181	0.060936	0.009064	0.190599	0.122568	0.075383	0.032493	0.268966	-0.023226
ifHate	0.030708	1.000000	-0.022540	0.009211	0.005510	0.038721	0.032632	0.022245	-0.002018	0.034562	0.033952
ifSarcasm	-0.049181	-0.022540	1.000000	0.018711	-0.002344	-0.088278	-0.066072	-0.081732	-0.041147	-0.100755	-0.046380
Public Holiday Or Not	0.060936	0.009211	0.018711	1.000000	-0.138603	-0.196269	-0.192009	-0.239374	-0.061951	-0.050437	-0.020257
Covid Cases	0.009064	0.005510	-0.002344	-0.138603	1.000000	0.433399	0.695948	0.424505	-0.116849	0.028474	-0.116312
Covid Deaths	0.190599	0.038721	-0.088278	-0.196269	0.433399	1.000000	0.799602	0.660437	0.022549	0.394204	0.193274
Covid Tests	0.122568	0.032632	-0.066072	-0.192009	0.695948	0.799602	1.000000	0.810640	-0.033373	0.325217	0.114351
Covid Vaccinations	0.075383	0.022245	-0.081732	-0.239374	0.424505	0.660437	0.810640	1.000000	-0.019049	0.329942	0.208065
female	0.032493	-0.002018	-0.041147	-0.061951	-0.116849	0.022549	-0.033373	-0.019049	1.000000	-0.234776	-0.092986
male	0.268966	0.034562	-0.100755	-0.050437	0.028474	0.394204	0.325217	0.329942	-0.234776	1.000000	-0.165771
others	-0.023226	0.033952	-0.046380	-0.020257	-0.116312	0.193274	0.114351	0.208065	-0.092986	-0.165771	1.000000



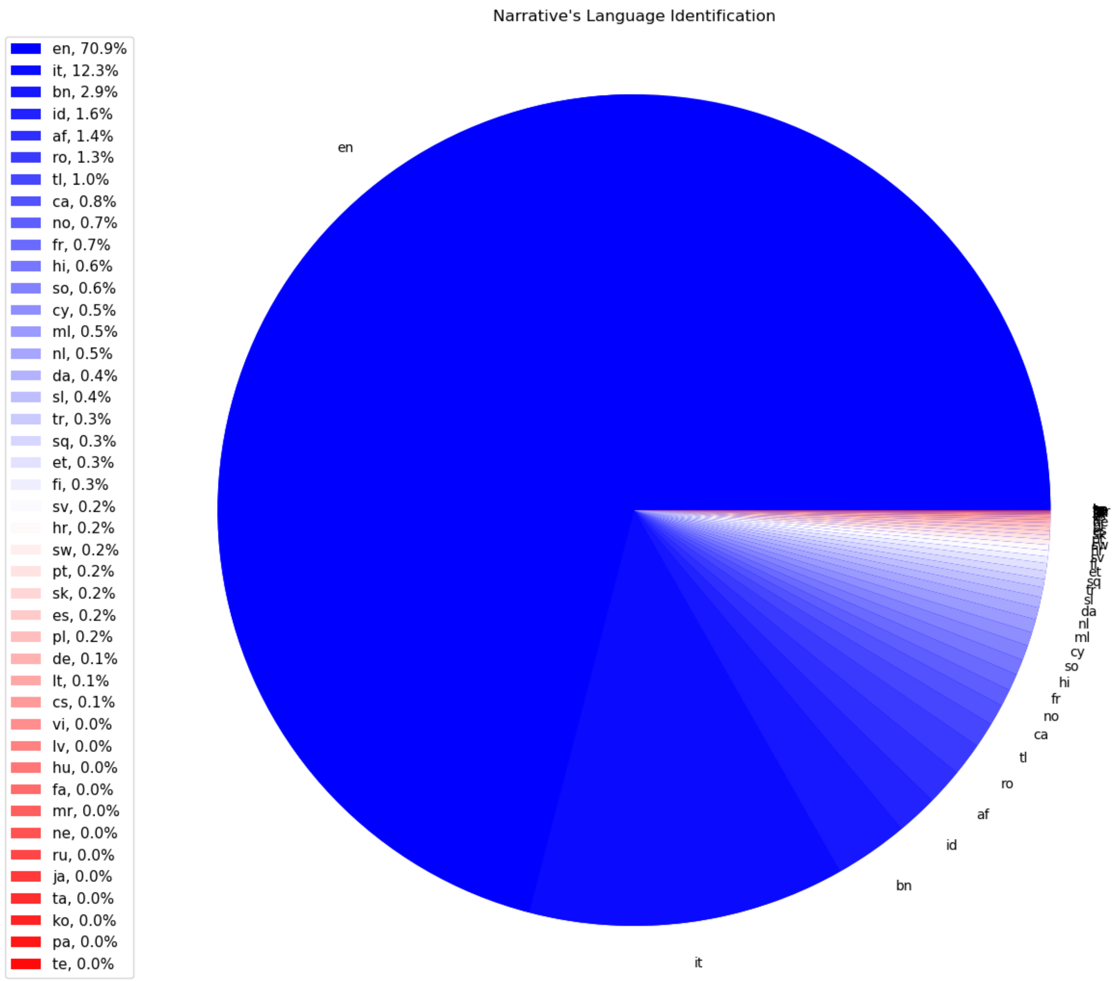
In addition, we think that users' interests may have an impact on whether or not they engage in hate speech. So we collected and ranked the top 15 interests of hate speech users and created the bar chart below.

As we can see, the number of "football" and "sports" in the top two is significantly higher than other interests that follow. We speculated that the users who are interested in these two topics may be more likely to produce hate speech because they are emotionally charged (both positive and negative) when watching football games or other sports games.



6. What insights do the "indirect" features you extracted tell us about the data?

When we investigate the relationship between users' interests and ages, the 15th top interest is Bollywood from the first figure in question 3. Unlike other top attractions, Bollywood implies a culture of objects by a particular group of users. Therefore, we want to determine the users' nationality and distribution by identifying the narratives of the posts in the PixStory Application from January 2020 to December 2022. From the figure below, the top three percent of languages used in the posts are English (70.9%), Italian (12.3%), and Bengali (2.9%) from the lists of ISO 639. The users' distribution might help the company decide on advertisement and business strategies.


Narrative's Language Identification

7. What clusters of users and/or posts made the most sense? Why?

Our analysis shows that males aged from 18 to 25 are much more likely to write negative posts and users with interest in football are mostly males.

From the first figure in this report, it is evident that there are more male users, and more negative posts are coming from males. Considering only the negative posts' numbers, males post 4.16 times that of females. In addition, females with ages over 30 are more likely to post negative. In contrast, the distributions of positive and negative posts by males' ages are similar.

Then, in question 4, it seems that 43.26% of hate speech comes from young adults (18-25). The second figure in question 4, implies that people aged 12 to 30 are more likely to hide their genders.
What's more, of all posts having hate speech, 43% of them come from males. At the same time, 35% of all hate-speech posts come from meals aged 12 to 30. This data emphasizes that young adults contribute most to writing hate speech and producing negative posts.

Moreover, from the second figure in question 3, there is a considerable large gap between the numbers of males' football interest and females' football interest. It might imply that male users will mark sports and football in their interests while females might only mark sports. Therefore, we can use this pattern as one support to identify users' genders for those users who hide their identity.

In summary, users' interest in football might be one of the practical ways to identify genders. Furthermore, males with ages from 18 to 25 years old are much likely to write negative posts. Therefore, this might be an excellent approach to classify users who produce negative posts.

Conclusion

After comparing and analyzing the data and features, we realized that young adult users, especially males, whose interests included football and sports, were more likely to produce hate speech posts. To create a better Web environment, we suggested the moderation mechanism of PixStory to be more prudent in checking posts that fit the above characteristics. Moreover, we found that some users' personal information was inaccurate or incomplete, such as age and gender, which caused certain obstacles during the data analysis. We recommend that PixStory would compulsorily request users for personal information without invasion of privacy.