**Simppl task**

## **Research Objectives**

1. **Examine, visualize, and identify patterns in media reliability scores from this list of unreliable news providers** [**https://iffy.news/disinfo-dashboard**](https://iffy.news/disinfo-dashboard/)
2. **Collect platform-specific engagement metrics including likes, shares, and the number of replies to each of the posts containing an article from these providers to understand user interaction with these posts.**
3. **Learn to tell a story with a graph, building intuitive and engaging data visualizations.**

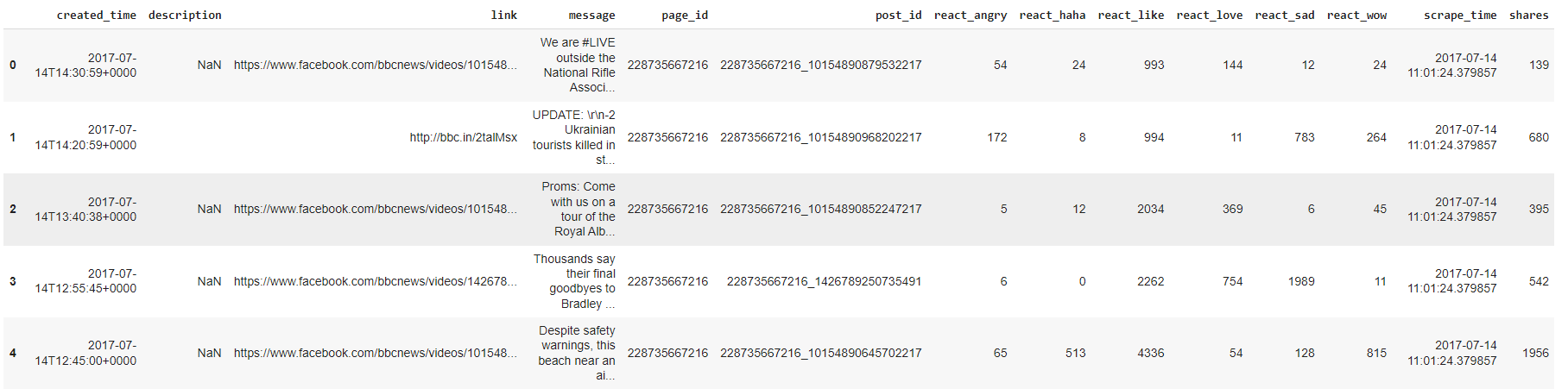
**Platform - Facebook**

After thorough research and analysis, 2 datasets of high relavance were selected for the required research. The datset mainly comprises of some reaction metrics of the users, post date & time, post content and shares. They have some interesting data of user interaction , reach, reacts , impressions which play an important role in identifying tredns and controversial posts which are likely to be detected as fake news or an upcoming trend. The dataset descriptions are as follows:

1st Datasets contains columns-

created\_time,description,link,message,page\_id,post\_id,react\_angry,react\_haha,react\_like,react\_love,react\_sad,react\_wow,scrape\_time,shares

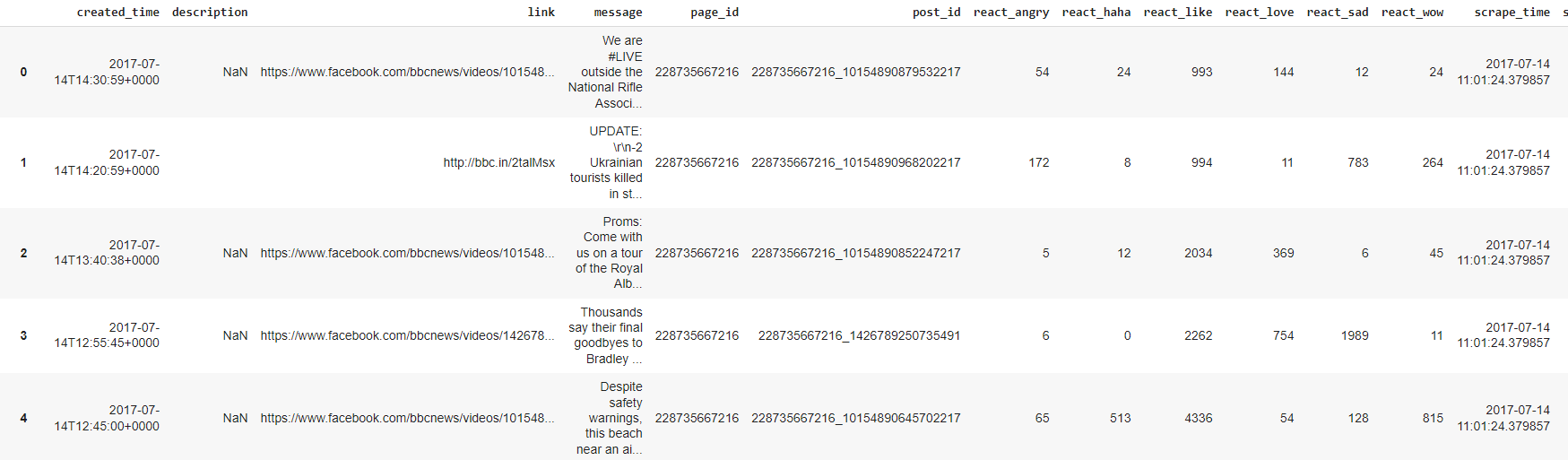
Size- 20k rows (shorter version of a sample which contains 100k rows) and 14 columns



2nd Datasets contains columns-

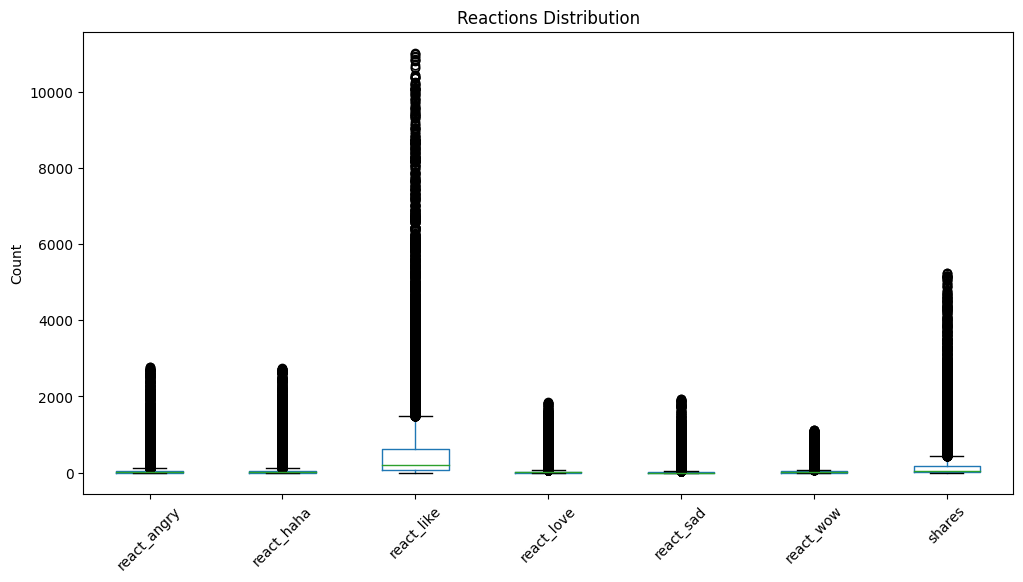
Page total likes;Type;Category;Post Month;Post Weekday;Post Hour;Paid;Lifetime Post Total Reach;Lifetime Post Total Impressions;Lifetime Engaged Users;Lifetime Post Consumers;Lifetime Post Consumptions;Lifetime Post Impressions by people who have liked your Page;Lifetime Post reach by people who like your Page;Lifetime People who have liked your Page and engaged with your post;comment;like;share;Total Interactions.

Size- 5000 rows, 19 columns

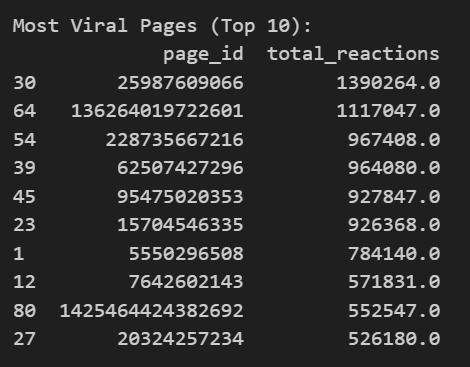


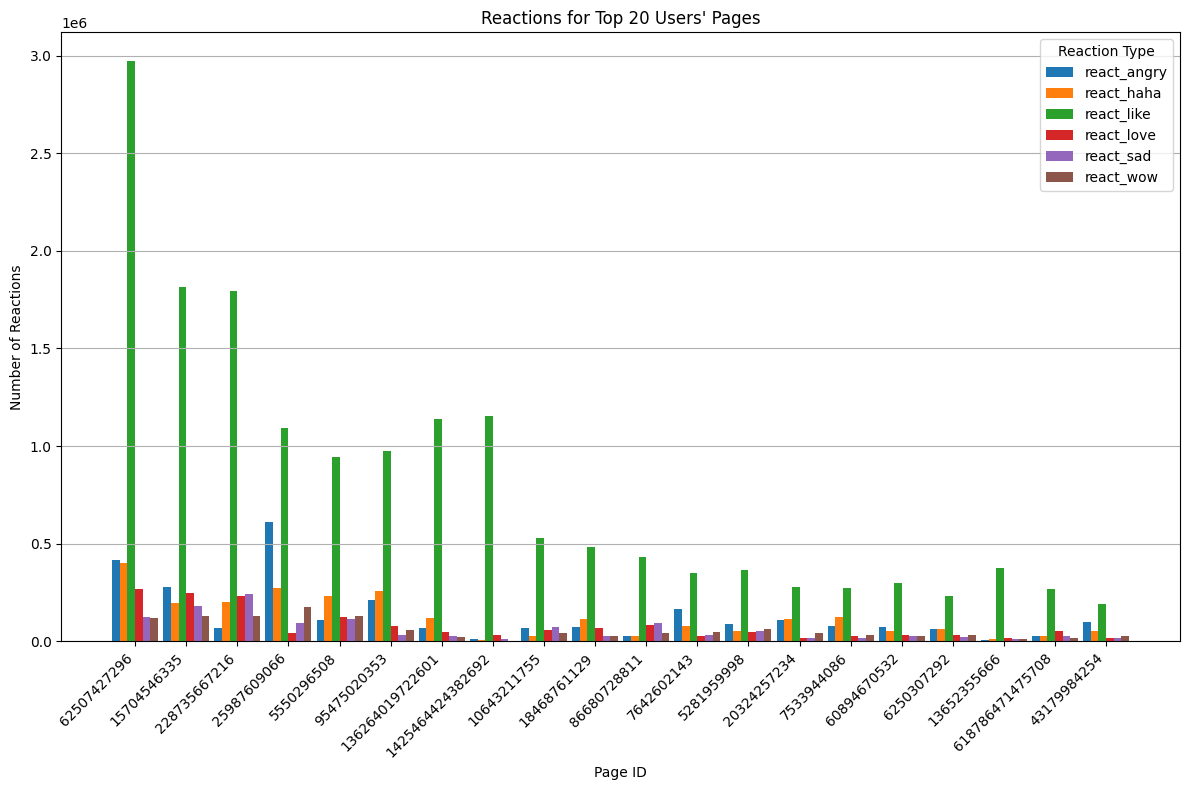
**Analysis of the first dataset**

The analysis takes place after cleaning of the dataset, which removes the outliers and null values and replaces them with the most probable values. The columns are kept intact as all of them are important for the analysis

1. Reaction distribution- 

2. Identifying the most viral pages based on the reactions- These page IDs are stored with another df containing the names of the pages. This is done for reducing the repetative entities



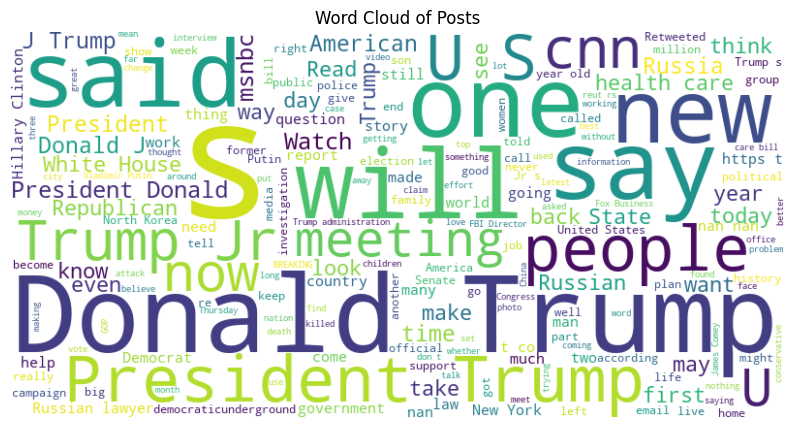
3. Identifying the dominating reactions of these posts 

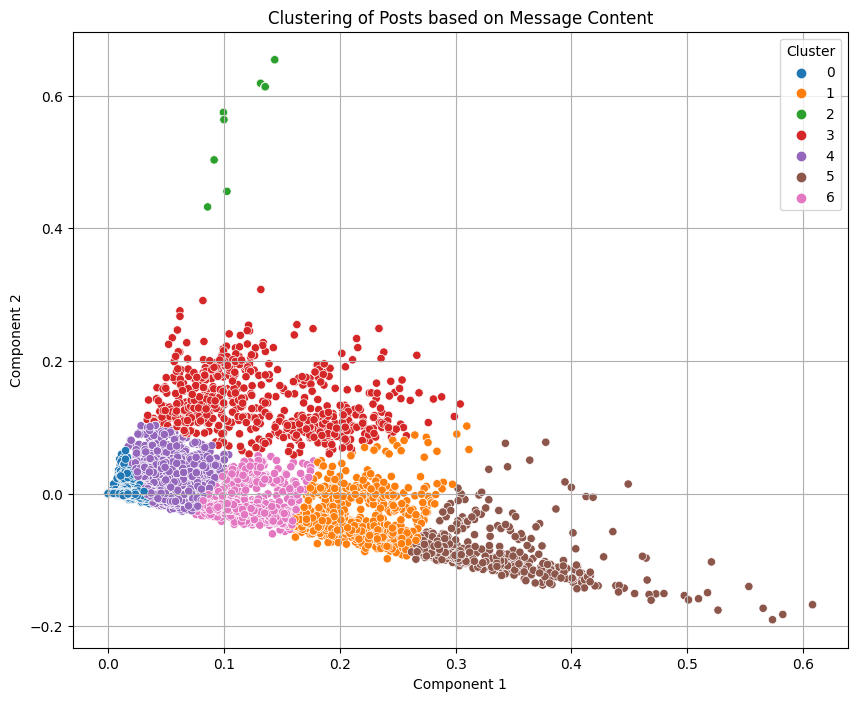
1. Identifying the content of the message

Text and Message preprocessing

Most Common Words:

[('trump', 3528), ('’', 3045), ('president', 2106), ('donald', 1578), ('“', 1481), ('”', 1411), ('us', 1244), ('new', 1038), ('one', 968), ('said', 962), ('russian', 796), ('people', 783), ('watch', 690), ('would', 683), ('house', 630), ('first', 627), ('meeting', 626), ('trumps', 617), ('news', 609), ('says', 577)]



1. Clustering the posts by their most important words

6. An insight on what the cluster contains.

Cluster 0 Texts:

1. The two brands will remain separate. "You will not be seeing 'Blue Man Group by Cirque du Soleil.'"

2. The Emmys got it mostly right!

3. Call it “Guaranteed lifetime coverage for all.”

4. Nearly half the states have refused to comply with a request to hand over substantial amounts of confidential and sensitive voter data.

5. The Democrat Party's become the largest hate group in this country.

Cluster 1 Texts:

1. President Donald J. Trump's suggestion that he would work with Russia was met with widespread skepticism.

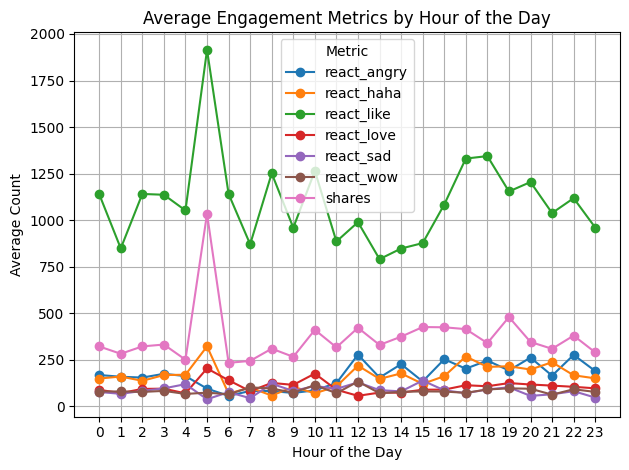
2. Democratic Rep. Gerry Connolly demanded that Vice President Mike Pence shut down President Trump's voter fraud commission's request for states to share private voter registration data, saying that it is "based on false claims made by President Trump and members of his administration."

3. Former late-night television show host David Letterman says that people need to “stop whining” about President Donald J. Trump and instead “figure out a way” to remove him from office.

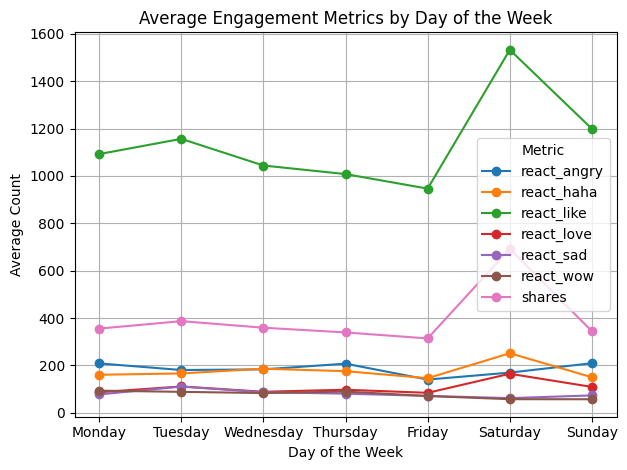
4. The right-wing Polish government is busing in a friendly crowd for President Donald Trump's visit.

5. President Donald Trump used to rag on Paris, but at a press conference during his visit to France Thursday he said that that city will be “very, very peaceful,” as they now have a “great president.”

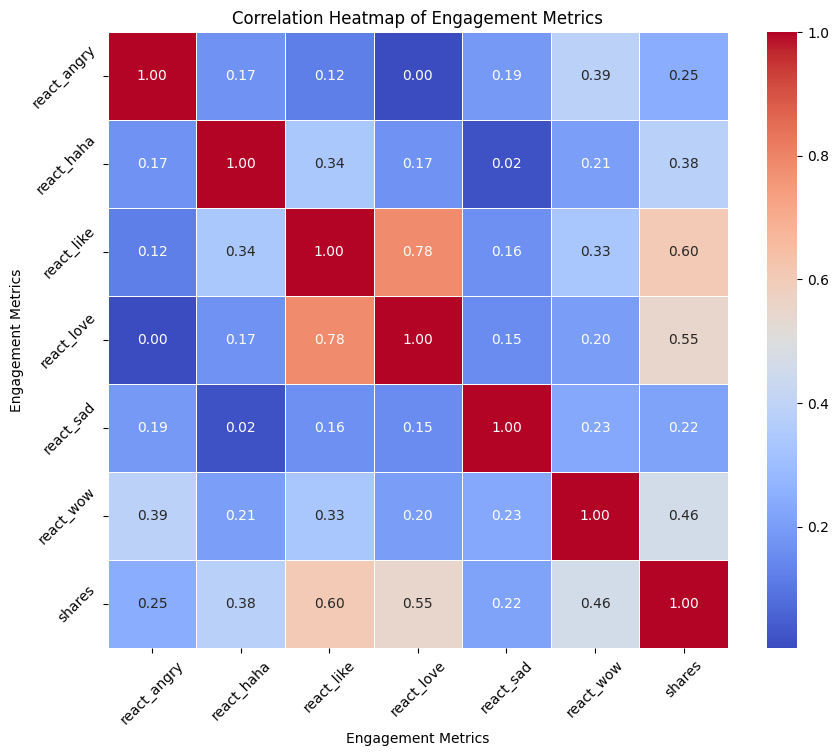
7. Identifying the most active days of the week based on the reactions- 5 am and nighttime seems to be the most activetimeline, but because the data is global , the timelines do differ



8. Identifying the most active days of the week based on the reactions- Saturday seems to be the most active days on facebook



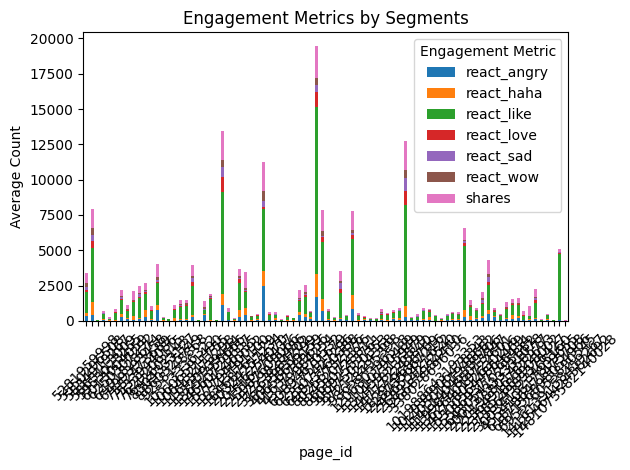
9. Finding the relation within the column- Some columns might be directly related while some might be Inversly related



We can notice that

* If a post is react with ‘react-love’ it is likely to be realted with ‘react-like’
* If a post is react with ‘react-like’ it is likely to be shared
* If a post is react with ‘react-love’ it is likely to be shared
* If a post is react with ‘react-angry’ it wont be in ‘react-like’ or ‘react-love’
* If a post is react with ‘react-sad it wont be in ‘react-haha’

10. Engagement metrics by reactions



Some of the most viral posts are either very controversial or very loved

Trump is left hanging after trying to shake hands with the Polish president's wife. \r\n\r\nMore on his keynote speech in Poland: bbc.in/2stMesr", '<https://www.facebook.com/bbcnews/videos/10154859051287217/>')

Post 344: ('"What an honor it is to pray for the President in the Oval Office." \r\n\r\nPHOTO: An intimate moment during a meeting between President Donald J. Trump and evangelical leaders shows them laying their hands on Trump\'s shoulders as his head is bent in prayer.<http://fxn.ws/2tg49m1>', '<https://www.facebook.com/FoxNews/photos/a.184044921335.134777.15704546335/10155656548031336/?type=3>')

Post 342: ('Musician Kid Rock has announced his bid for the U.S. Senate in Michigan.\r\n\r\nRepublicans in Michigan have reportedly been considering Rock to run on the GOP ticket in 2018 to challenge incumbent Democrat Debbie Stabenow for her seat.<http://fxn.ws/2ujYTmt>', '<https://www.facebook.com/FoxNews/photos/a.184044921335.134777.15704546335/10155656751086336/?type=3>')

Post 17815: ('"I think this is a teachable moment, not just to you but to the people who are watching." - Ice Cube', '<https://www.facebook.com/Maher/videos/10154783827827297/>')

Post 143: ("You'll never guess when this photo was taken...\r\n\r\nThis woman is adding colour to black and white photos from history.", '<https://www.facebook.com/bbcnews/videos/10154875262277217/>')

Post 16211: ('Seven years ago yesterday (March 1st), I married Douglas Brunt. Three kids, a few wrinkles, and a lot of laughs later, he is still the best (grown-up) person I know.', '<https://www.facebook.com/MegynKelly/photos/a.1441271546135313.1073741829.1425464424382692/1539608909634909/?type=3>')

Post 492: ("Viral Video Alert! Last night, before boarding Marine One, President Donald J. Trump retrieved a U.S. Marine's hat, and the moment is taking the Internet by storm.<http://bit.ly/2t2yhkN>", '<https://www.facebook.com/FoxNews/videos/10155640699266336/>')

Post 394: ('A military supporter played "Taps" Tuesday afternoon to honor the 16 service members killed in Monday\'s plane crash.<http://fxn.ws/2taoLfF>', '<https://www.facebook.com/FoxNews/videos/10155651544106336/>')

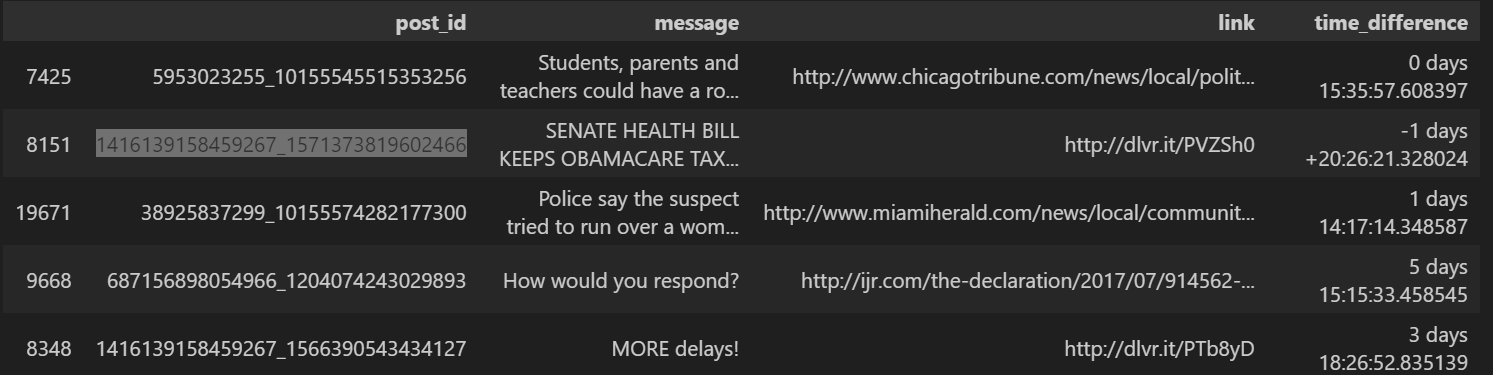
Post 200: ("'You could ask my dad Stephen Hawking any question'\r\n\r\nWhat it was like to grow up as the daughter of the famous scientist.", '<https://www.facebook.com/bbcnews/videos/10154866313777217/>')

Post 158: ('A disabled fan got the most incredible Coldplay concert experience ever. [http://bbc.in/2sVPbX6\r\n\r\n(Via](http://bbc.in/2sVPbX6/r/n/r/n(Via) BBC Entertainment News)', '<https://www.facebook.com/bbcnews/videos/1412464658837743/>')

11. Time series analysis

* How long did a post take to get famous. This can be identified by checked the total reactions within the ‘post\_creation\_time - scraping\_time’

Some posts got the most ‘react-angry’ in the least time

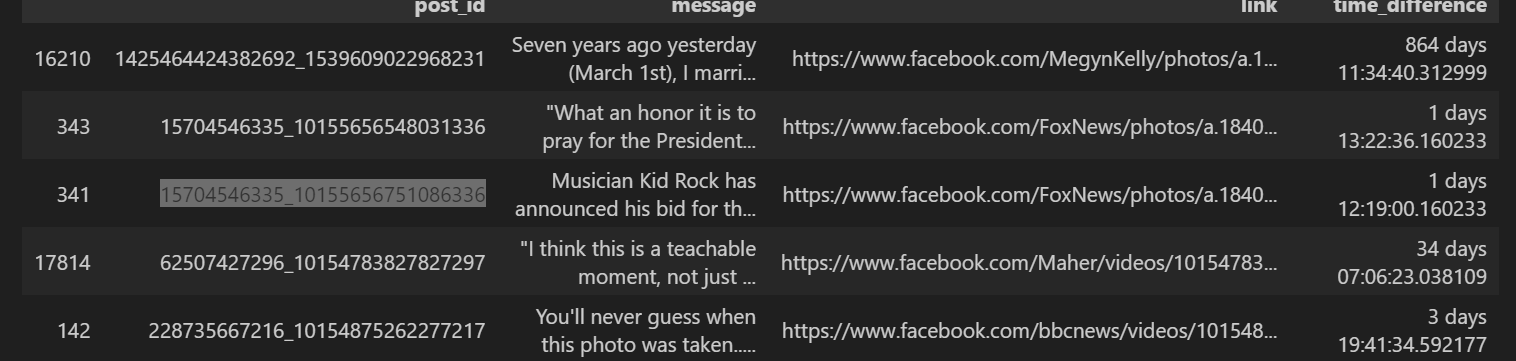


8851-<https://thefederalist.com/2017/07/14/5-backroom-deals-inside-latest-senate-health-care-bill/?utm_source=dlvr.it&utm_medium=facebook>

19671<http://www.miamiherald.com/news/local/community/broward/article160997344.html>

9968<http://ijr.com/the-declaration/2017/07/914562-terrifying-video-shows-moment-defenseless-woman-kidnapped/?utm_source=Facebook&utm_medium=Owned&utm_term=ijpolitics&utm_campaign=ods&utm_content=Politics>

Some posts got the most ‘react-like in the least time



16210-<https://www.facebook.com/MegynKelly/photos/a.1435328533396281.1073741828.1425464424382692/1539682359627564/?type=3>

343-<https://www.facebook.com/FoxNews/photos/a.184044921335.134777.15704546335/10155656548031336/?type=3>

341-<https://www.facebook.com/FoxNews/photos/a.184044921335.134777.15704546335/10155656751086336/?type=3>

12. Estimation model comparison

The estimation of how likely the reaction will turn out to be famous based on the current reactions. The metrics of all the reactions - like, love,sad, angry, wow, haha are all trained on all different models and compared. As we can see Linear regression, Ridge and Lasso Regression models give the best analysis and Random Forest and Gradient boosting gives the worst prediction.

TRYING TO PREDICT THE POSSIBILITY OF A POST TO GET FAMOUS BY CALCULATING THE CURRENT REACH AND LIKES. WE WILL NOW SELECT THE BEST MODEL FOR THE PREDICTION.

Linear Regression Mean Squared Error: 4.823806201840176e-23

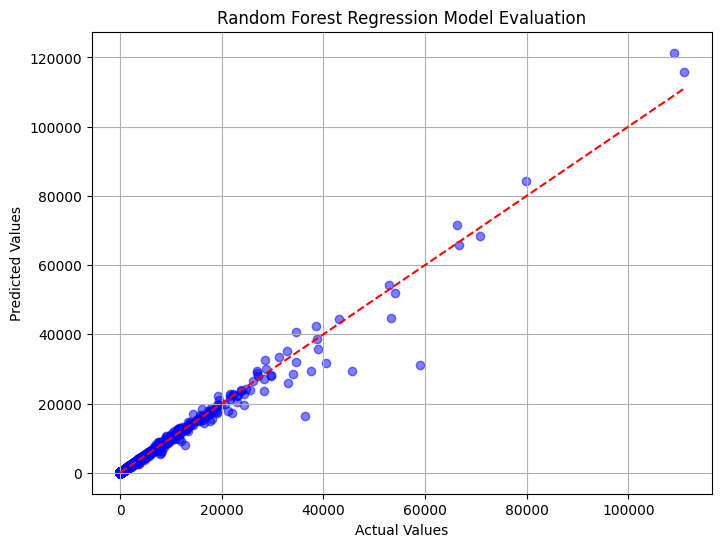
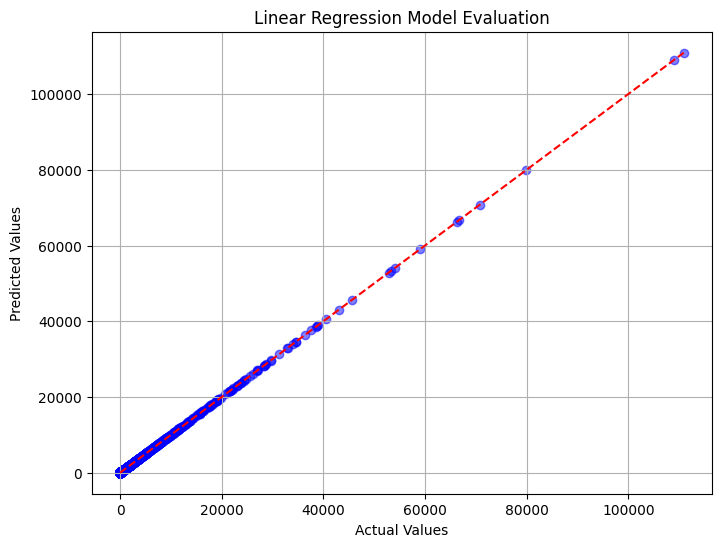
Random Forest Regression Mean Squared Error: 596012.8706160705

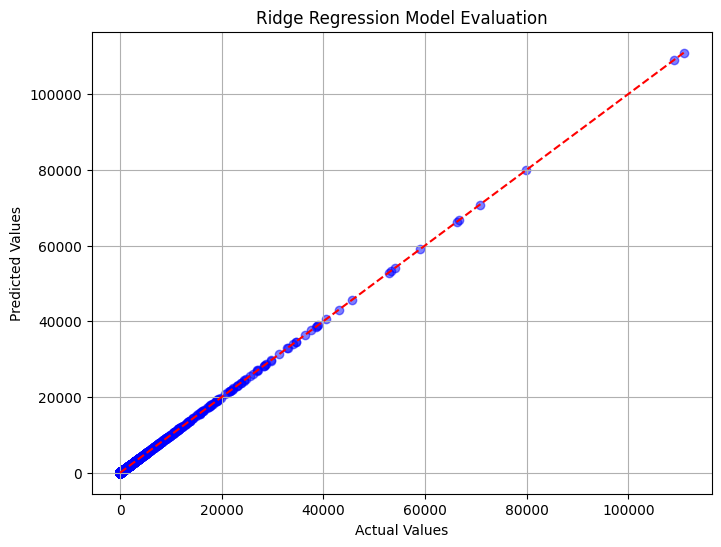
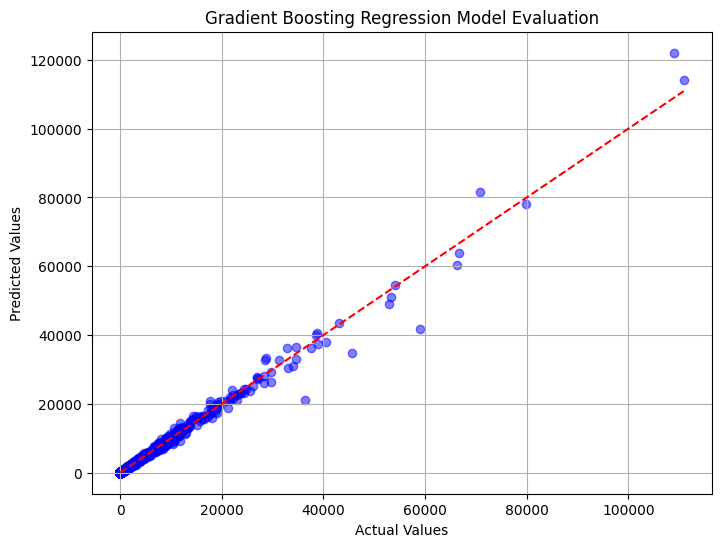
Gradient Boosting Regression Mean Squared Error: 336916.1036886329

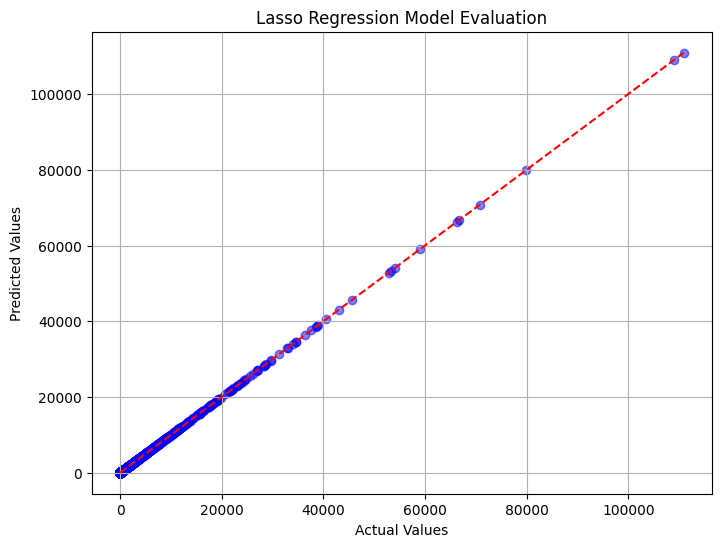
Ridge Regression Mean Squared Error: 1.0678233055749774e-13

Lasso Regression Mean Squared Error: 0.0072616223438031246

THE ABOVE EVALUATION TELLS US THAT LASSO,RIDGE AND LINEAR REGRESSION MODELS ARE THE BEST FOR PREDICTING THE TOTAL REACTION.



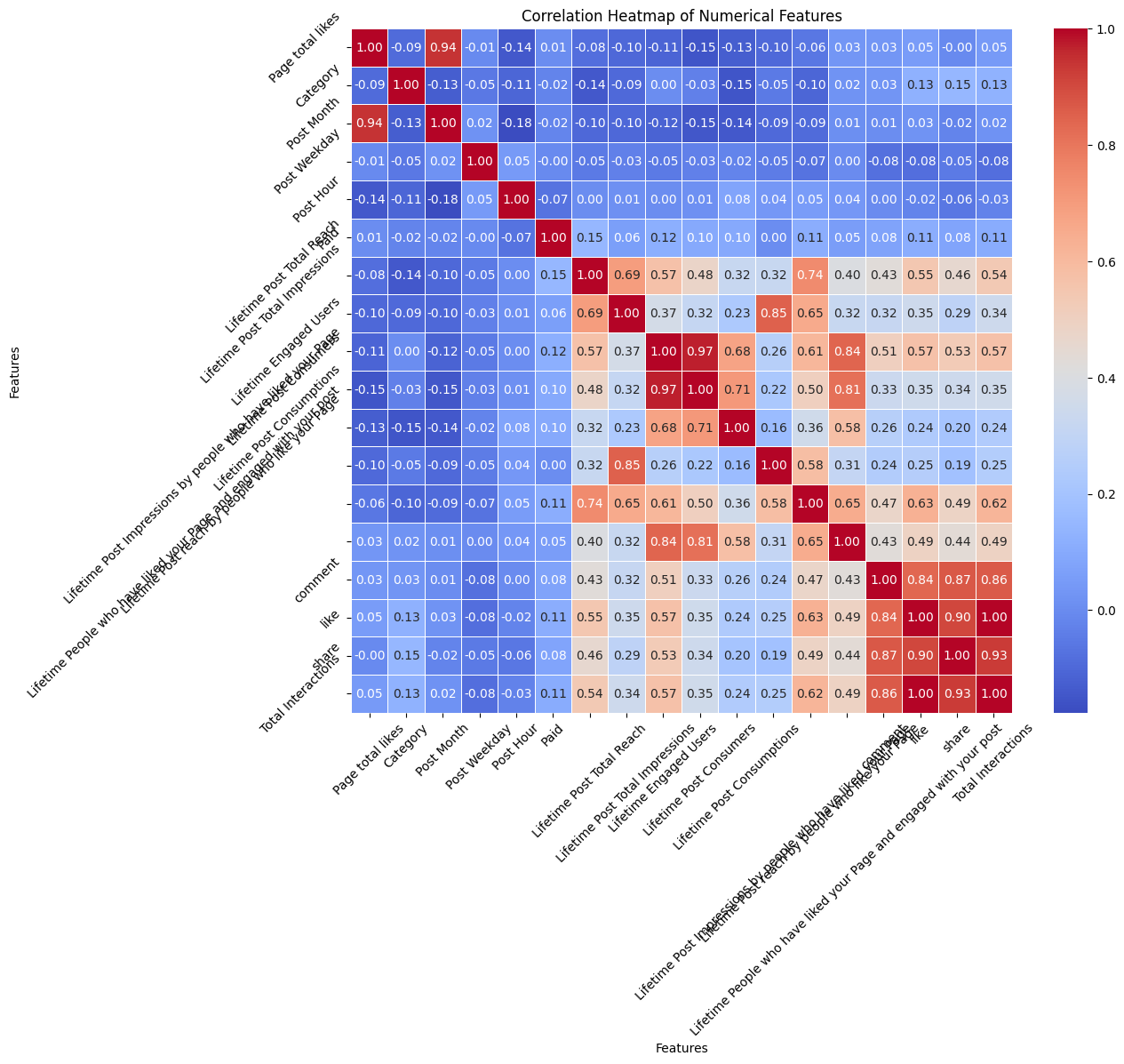




2nd dataset

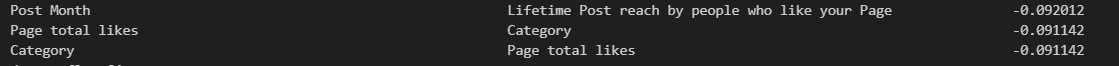
The second dataset comprises of some unique features like post impression, Likes, reacts, post consumption, consumers, category,comments, likes,shares and total impression.

The correlation of these features are as follows:

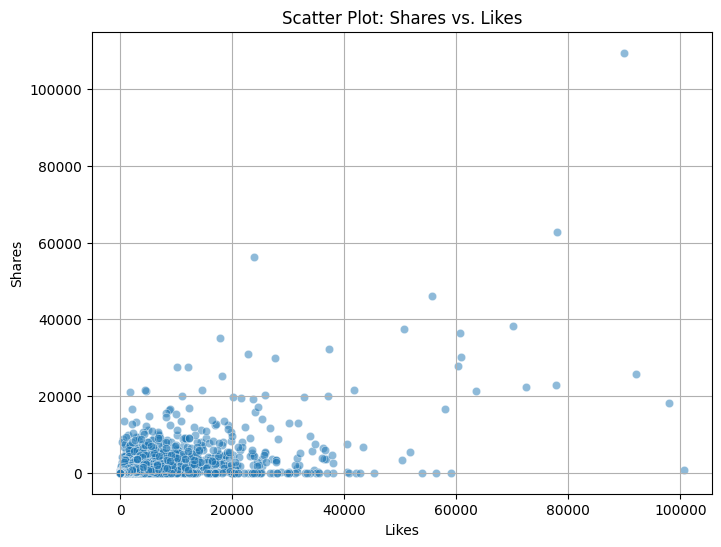


Most related features with their similarity score

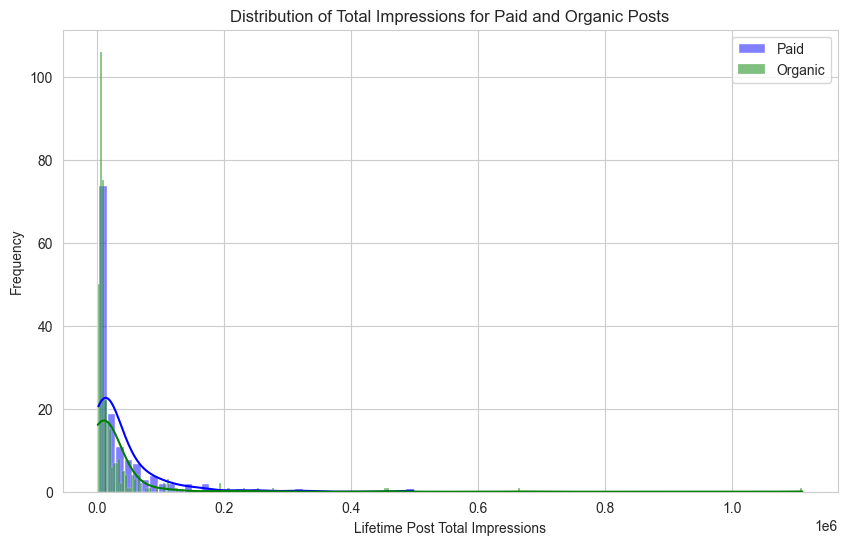
Most unrelated features



Checking the metrics -shares and likes and their relation



Checking the metrics -shares and likes and their relation



Hence the data analysis is done on 2 extensive and highly coveted dataset. Another analysis of weighted graph network could be make if the data of the friends , followers were present. It would take some computational resources, and research for the same. The analysis could give us useful insights of what is popular in which friend group, and the same pots could be recommended to another group of friends.

I tried implementing the same but dataset content columns were not adequate.

I tried webscraping one of the website mentioned in the assignment task - BitChute. It lies very close to Simppl’s goal in my opinion which identifies fake news and genuine news along with user metrics.

The process took a little more time than I anticipated, so i went ahead with the facebook dataset after the cumbersome task of acquiring it among the security restrictions of Meta.