

CS 6350

ASSIGNMENT ____3_____

Names of students in your group:

Nimrat Bedi nxb200004

Rishab Bansal rxb190055

Number of free late days used: ____1_____

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

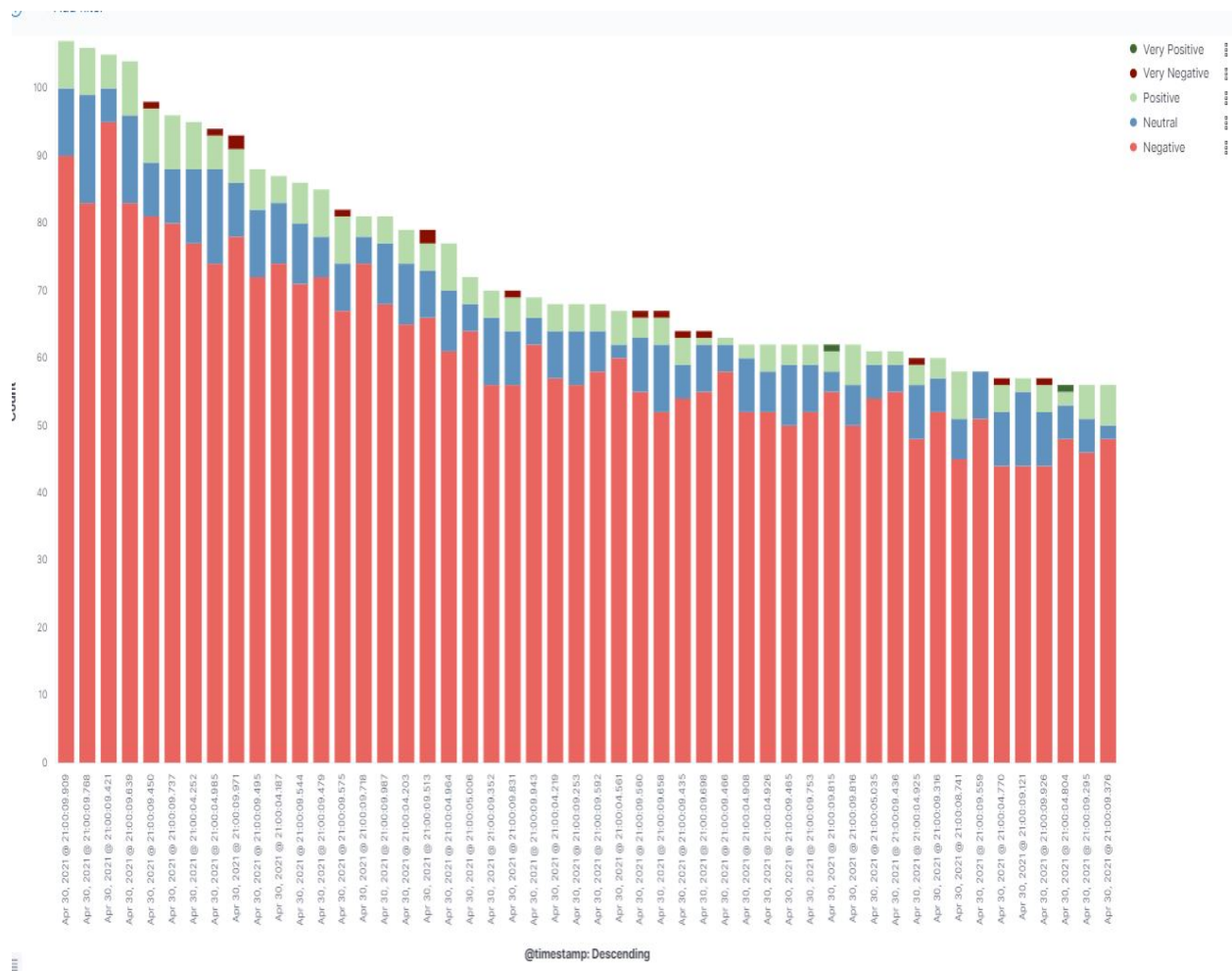
Please list clearly all the sources/references that you have used in this assignment.

Part 1

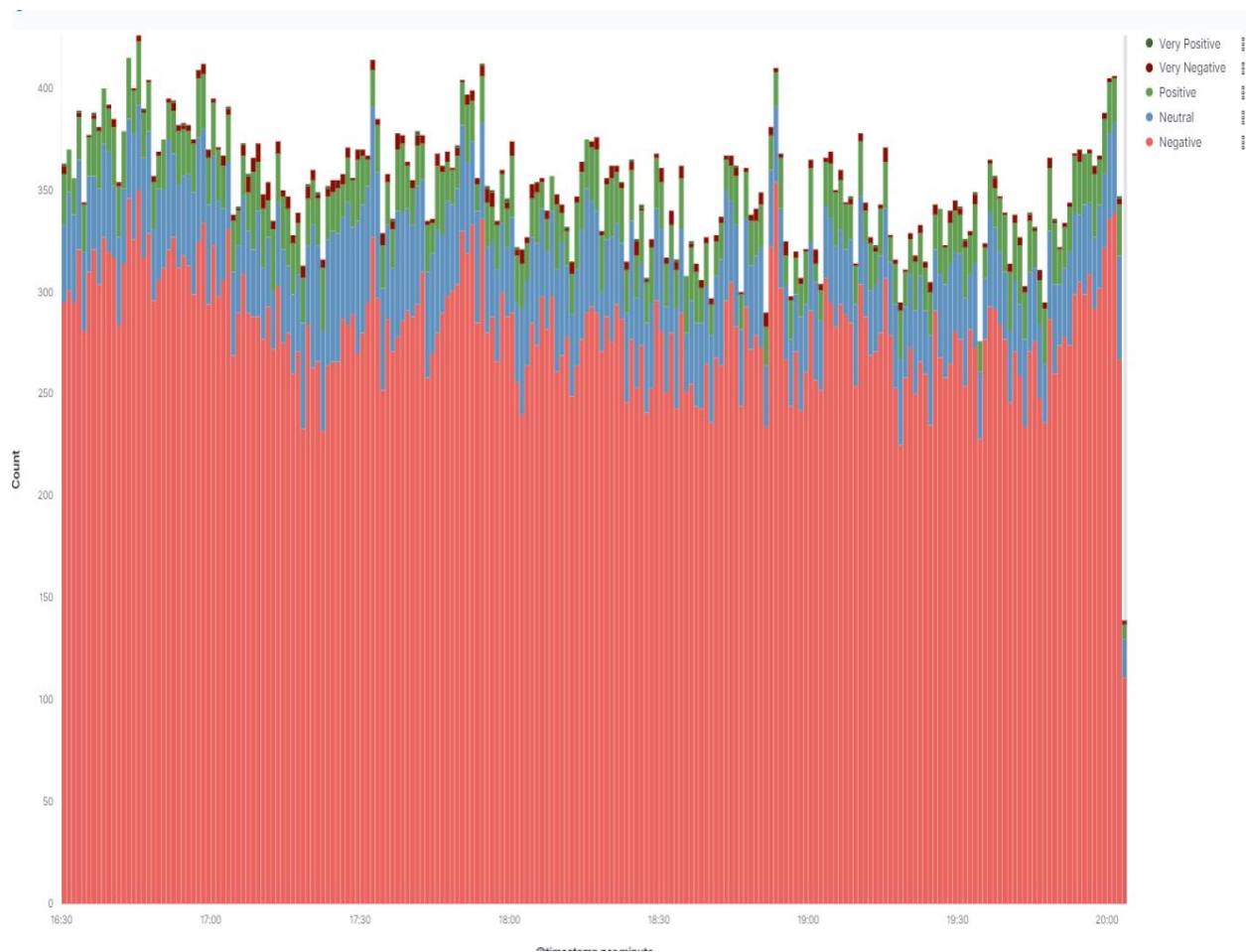
Summary Results

We have streamed Twitter Data with keyword “India”. We can observe the per minute sample sentiment split graphs below. We observed the data on 30th April over a period of 2 hours and 1st May over a period of 4 hours

30th April data:



1st May data:

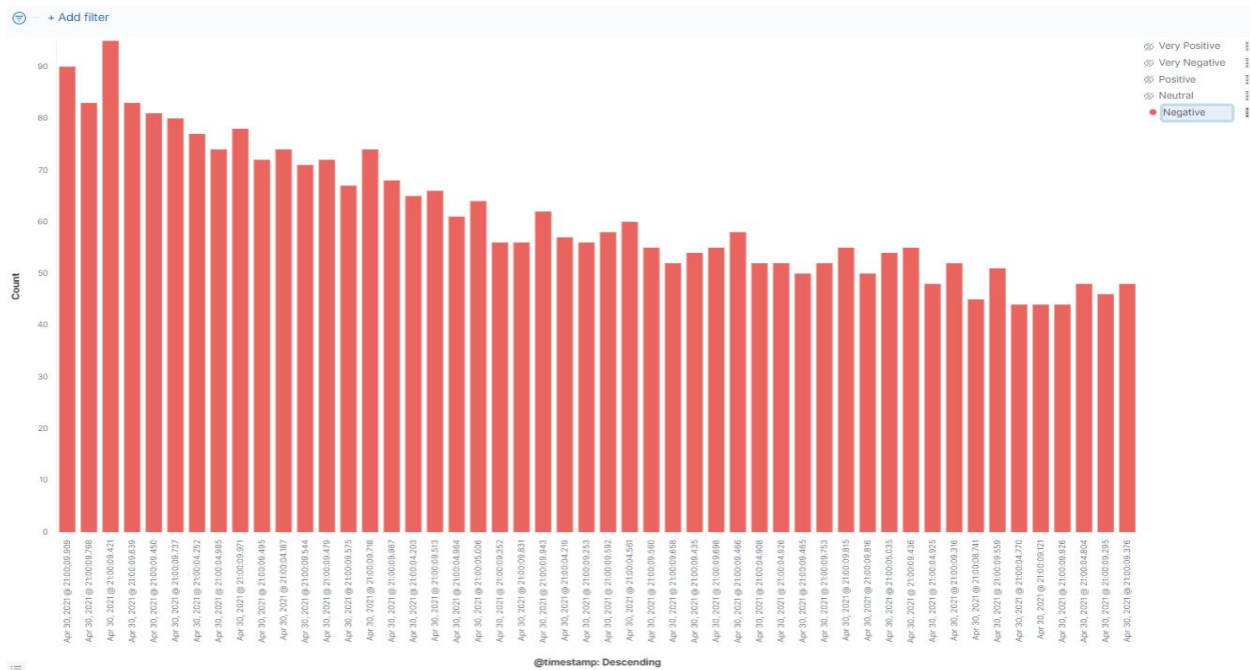


Insights

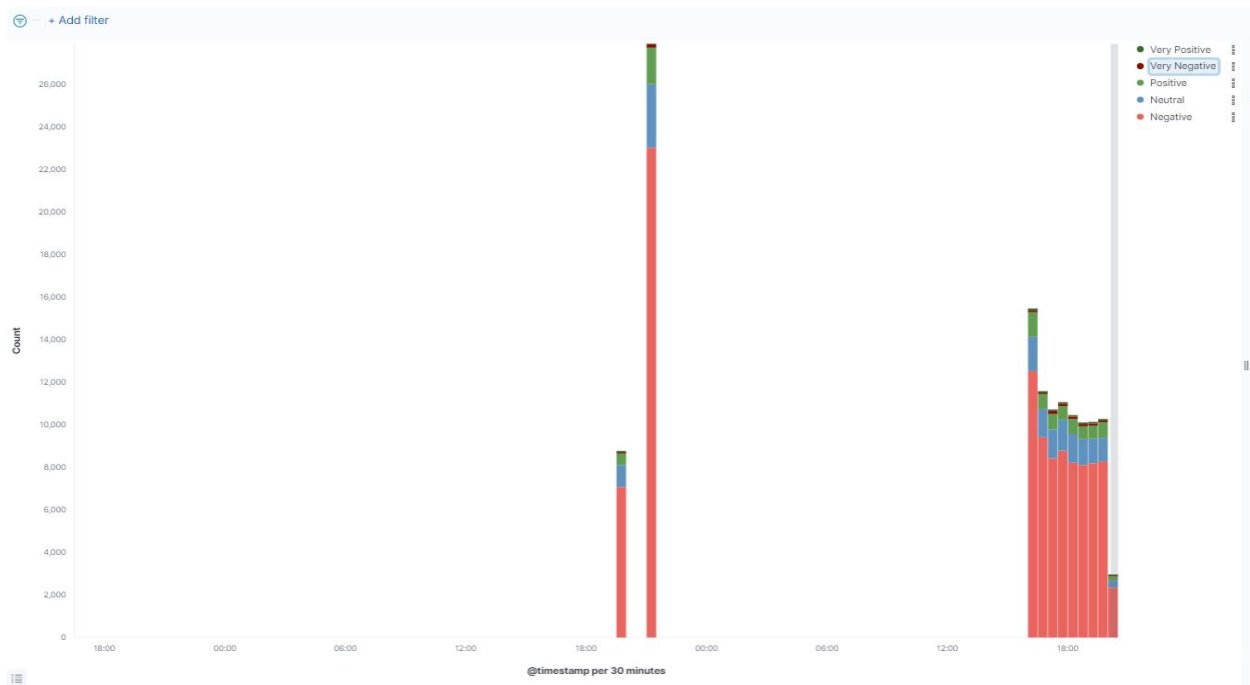
Here we can see, as right now in India due to covid'19 - situations are really bad so most of the tweets are negative.

Conclusion: From both the days we visualized the data: mostly negative tweets are coming under "India" topic.

As you can see from the image below the red bar chart denotes negative tweets.



Two day data representation: (30th May and 1st May)



Part 2

Summary Results

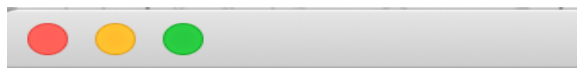
Dataset Information:

We took the Twitter dataset from SNAP repository.
Description of Dataset: Social circles from Twitter

Insights:

Number of Nodes and edges can be found by counting the vertices(vertices.count()) and edges (edges.count()) to cross check if the graph was created properly, doing so resulted with 81306 vertices and 1768149 edges which is equal to the Dataset Information given.

1. Outdegrees in descending order specifies which 5 user ids cast Highest count of tweets for others. User 3359851 made 1658 tweets followed by 59804598 with 1295 tweets and then user 5442012 with 911 tweets followed by 15102849 with 883 tweets and user 88323281 with 867 tweets. These 5 people can be considered as frequent participants in the Twitter.



```
|[3359851,1658]  
|[59804598,1295]  
|[5442012,911]  
|[15102849,883]  
|[88323281,867]
```

2. Indegrees in descending order specifies which 5 user ids got Highest number of tweets from others. User 40981798 got 3543 votes followed by 43003845 who got 3101 tweets nodes and then user 22462180 got 3092 followed by 34428380 who got 3070 tweet nodes and user 115485051 with 2523 tweet nodes. These 5 people can be rated as authentic users as they got most tweets from others.



```
[40981798,3543]  
[43003845,3101]  
[22462180,3092]  
[34428380,3070]  
[115485051,2523]
```

3. PageRank is Highest for user 115485051 followed by 505.46 like the indegrees result and 116485573 in 4th place specify that these are popular users with most of the in links from other existing admins and other tweets. User 813286 at 3rd and user 7861312 at 5th are outliers who could have got tweets from users and admins with higher pagerank even if they got a smaller number of votes than other 3 resulting in a better page rank in comparison to Indegrees result.



```
[115485051,505.459331233658]  
[116485573,438.79665455196135]  
[813286,128.21304006757174]  
[11348282,124.91393287517268]  
[7861312,106.21577943335762]
```

4. Largest connected component have id 0 and have 40185 nodes which is equal to number of users participated on twitter and remaining components are with a smaller number(3&2) of nodes and can be said as weak components with less number of participants among them connectivity wise.



```
[0,40185]  
[315,89]  
[843,34]  
[234,32]  
[2924,17]
```

5. Users 40981798,22462180, 34428380 is in top 5 for trainglecounts.



[40981798,50102]

[22462180,45723]

[34428380,45388]

[43003845,45382]

[3359851,31062]