SHIKHAR TIWARI

I have Analysed data of four different social media sites like Twitter, Reddit, Parler,and BrightKite(not functioning now, but data was available).

Twitter

Data Collection:

pkl file ("DcpNorthdelhiTweets.pkl") has been uploaded. Code is also present in the ipynb file.

```
In [3]: #tweepy api
        auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
        auth.set_access_token(access_key, access_secret)
        api = tweepy.API(auth,wait_on_rate_limit=True,wait_on_rate_limit_notify=True)
```

```
In [ ]: #Data collection
        tweets=[]
        for status in tweepy.Cursor(api.user_timeline,id='DcpNorthDelhi',tweet_mode='extended').items():
            tweets.append(status)
```

Table of PII's collected

Table 1

	PII_types	count
0	emails	0
1	mobile_numbers	6
2	aadhar_card_numbers	0
3	pincode	12
4	vehicle_numbers	8

Table 2

	Tweet ID	no_of_emails	no_of_mobile_numbers	no_of_aadhar_card_no	no_of_pincodes	no_of_vehicle_numbers
6	1361152031545692163	0	0	0	1	0
14	1371764534826790918	0	1	0	0	0
30	1380546760079118336	0	0	0	1	0
36	1378975620038725634	0	1	0	0	0
55	1318548815188795394	0	0	0	0	1
57	1318421790574604288	0	0	0	0	1
63	1323518100780691456	0	0	0	0	1
91	1328745540129996800	0	1	0	0	0
114	1312552935138586624	0	0	0	0	1
117	1374672241439338497	0	1	0	0	0
129	1347780328266424321	0	0	0	1	0
142	1329728857159524361	0	0	0	1	0
182	1317163517330427904	0	0	0	1	0
183	1346734208023101440	0	0	0	2	0
200	1381103964985253888	0	0	0	1	0
236	1343951715872120832	0	0	0	1	0
273	1312575803675164672	0	0	0	0	1
279	1338318924769705986	0	0	0	0	1
287	1317400993760387072	0	0	0	1	0
299	1318876851054366722	0	0	0	0	2

Examples:

@DelhiPolice @CPDelhi Dear Sir, My car White Suzuki Vitara Breeza has been stolen from Shakti Nagar, D elhi-110007 on yesterday 14/02/2021 @ 8 am. In this regards, when i visited my nearest PS Roop Nagar, t hey told me that we cant register your FIR as DP website is not working.

Delhi police please help me 8700034868 https://t.co/6afOVQwYIk @DcpNorthDelhi @LtGovDelhi @CPDelhi @DelhiPolice @CPCB_OFFICIAL @PMOIndia @HMOIndia कश्मी री गेट थाने के अंतगगत कश्मीरीगेट मैटरोस्टेशन गेट न०2 के सामने up न० के टाटा 407 टरक सवारी व सामान की सुबह 1 0:00 बजे से रात 11:00 बजे तक सवारी व सामान की ओवरलोड ंग करते है इन्हें हटाए! https://t.co/TpaFvyosj Ek police karmchari 500 rs mask ~~challan katne m apni jaan laga deta h kabi koi Bina paise diye na Chala Ja ye~~

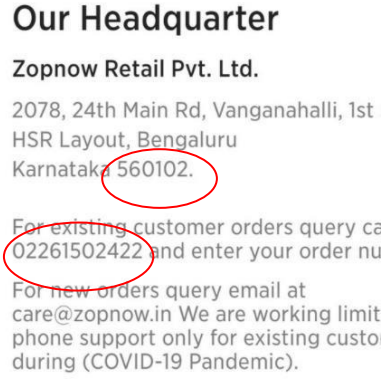
Lekin Meri car DL10CS3196 chori hue 4 months ho gye uske liye effort ni dikhane jabki chor ki footage bi h

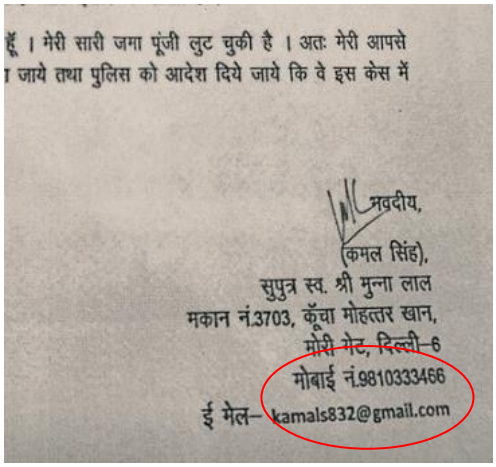
@DcpNorthDelhi @LtGovDelhi @CPDelhi @DelhiPolice My Bike was theft bullet in RED Colour from Durg a Park Sagar Pur in 01 Nov Morning 04:20 Am and bike No. DL9SBK8597 FIR No. 028328.

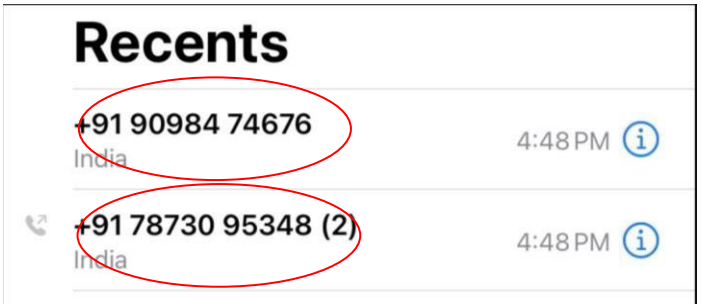
@DcpNorthDelhi Dear sir I am from Mumbai and I have paid online to a dealer in Sadar bazaar Delhi arou nd 50000/- to supply goods. Now he is not receiving call.Shinning rubber ~~chemicals (Gourav Sheth) 1321, PanMandi,SadarBazar, Daulat Ram Marg, Narain Market,~~ Sadar Bazaar, 110006

```
Number of tweets with media:  142
```

Out of these tweets, I found PII in some tweets.







Response times by police department in replying to tweets.

a)

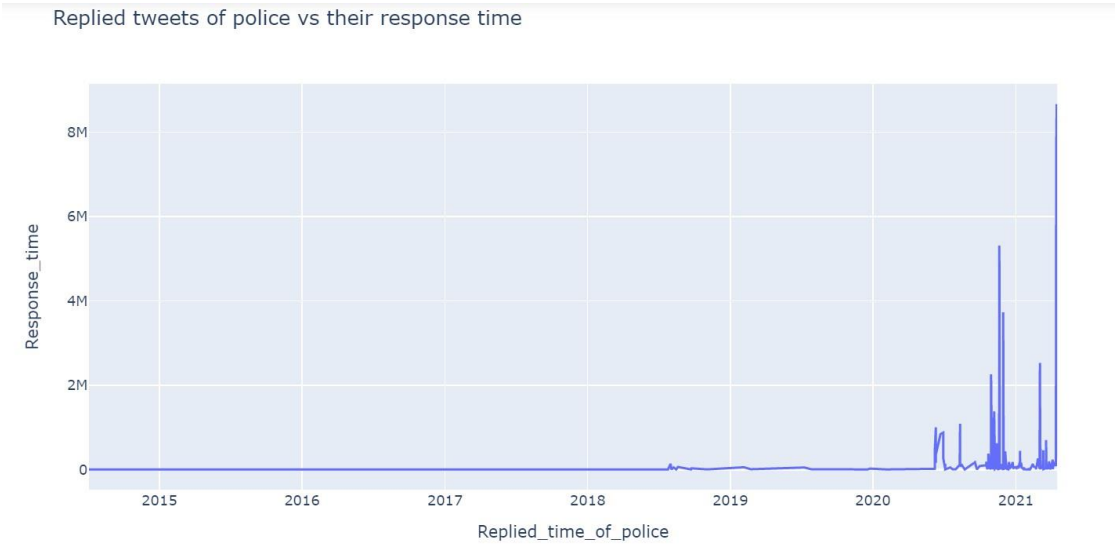
```
mean_response_time: 3 days, 13:55:56.583333
minimum_response_time: 0:00:00
maximum_response_time: 100 days, 6:53:46
standard_deviation_of_response_time: 11 days, 14:08:22.663235
```

- Minimum response time is 0 sec, which shows that some tweets are getting replied by bots.
- Maybe police department uses some bot for basic queries.
- Maximum response time is almost 100 days, this shows that a human has responded to this query.
- Standard deviation, mean also shows that this account is majorly handled by humans and not bots.

Inferences:

- High maximum response time, standard deviation shows that this account is majorly being handled by humans.
- Few quick responses close to zero seconds shows some bots replying to some basic/ common queries.
- Some queries might have been missed by police, which indicates such high response times.
- Police should work on their response time as a mean response time of 3 days is very high.
- A separate cell to cater to these needs should be made.

b)



Inferences:

- The account has not been very much active before May, 2020.
- After May,2020 a lot of activity is registered. Maybe because in may 2020, covid became a serious menace, people started communicating with police on twitter handles and police also started using twitter more seriously.
- There are many peaks (April 2021, March 2021, Nov 2020, October 2020).

- Some tweets have been replied pretty fast, while some tweets have been left un answered for long. Maybe because, department is getting more queries at certain time than other times.
- Some tweets might not have needed details, as a result police have not responded to them.
- We can see that police is not responding to users after 17:00 hrs. (5' o clock evening) till 2:00 hrs. (2 am in the morning). And responding unevenly during other parts of the day.
- This is possible because maybe police department cell which looks into social handles closes after 5 o clock.

Classifying tweets into categories

Table showing classification of tweets in different categories:

	Tweet_id	Tweet	classification
0	1322051191929384960	@PMOIndia https://t.co/mFC7AY5FAy	other
1	1321004437951819778	@iimc_majhi @DcpNorthDelhi for necessary action	other
2	1372067990812962818	@thehemanttweet @Live_Hindustan @DcpNorthDelhi...	report
3	1318382785086763010	ये है तिमारपुर दिल्ली थाने के सिपाही । मोबाइल ...	accountability
4	1291586896032706561	#IndependenceDayQuiz \nToday's question\n\nWha...	other
5	1322517197604859908	Can we wake up now?\nAll the sleeping official...	acoountability
6	1356930762042265600	Scenes from our so called plastic free Roshnar...	report
7	1335811345975889920	@DcpNorthDelhi @CPDelhi @AmitShahOffice @AmitS...	accountability
8	1317425489925976069	In BUARAI, BUARAI CHOWK to SATY VIHAR GATE dai...	request
9	1327895274212052992	@DcpNorthDelhi @CPDelhi @LtGovDelhi @DelhiPoli...	request
10	1319106698385051648	@GaganGarg711 @DcpNorthDelhi @DCPEastDelhi @DC...	other
11	1234132576861704192	अवंतिका मार्केट में दंगे हो रहे है क्या @Delhi...	other
12	1209558227370659840	They're evacuating Mukherjee Nagar in Delhi be...	accountability
13	1209632071653543936	@DcpNorthDelhi why they asked to leave PG and ...	request
14	1340017977832263681	Please Help me sir.	request
15	1333072526386155521	@CPDelhi 18 सितंबर की देर रात 5 गुंडे मिलकर 1 ...	accountability
16	1229429479732170753	" अभिमान मेरा है दिल्ली पुलिस "\nThe Delhi Pol...	appreciation
17	1029733936430075905	Today, On Independence Day Celebrations.. led ...	appreciation
18	1176794633570963456	@DcpNorthDelhi @LtGovDelhi @DelhiPolice What a...	report
19	1058405376981434368	Briefing of ACPs and SHOs of North District Po...	appreciation
20	1331659797360361472	@DcpNorthDelhi Sir From last three months ja...	report
21	1328745540129996800	@LtGovDelhi @DcpNorthDelhi @CPDelhi @PMOIndia ...	report
22	1265255528055390208	@DcpNorthDelhi mam i am a upsc aspirant from f...	request
23	1092455845005312000	Brave constable Raj Kamal Meena of PS Khyala s...	appreciation
24	1340908730196508673	@DcpNorthDelhi Dear sir I am from Mumbai and I...	report
25	1334359299938664449	@rishirajshanker @DcpNorthDelhi	other
26	1333660450320367618	बुरारी के अंदर वेस्ट कमल विहार में 6 लाइटों वा...	report
27	1367485856584450054	वजीराबाद के दिल्ली जल बोर्ड के स्टाफ क्वार्टर ...	report
28	1325043111031164928	लगा तार बढ़ रही है संगम विहार वजीराबाद में चोर...	report
29	1379100010969714688	and is good for nothing. We have all the evide...	request

Classification Number of tweets

Report : 9

Request : 6

Appreciation : 4

Accountability : 5

Other : 6

a) Categorizing different types of tweets.

Different categories can help police in the analysis of tweets in a much better way.

Request:

- Tweets having imperative tone, looking for help can be put in this category
- These tweets should be responded without much delay.
- Words like “help”, “please”, “kindly” can be used to identify requests.
- These type of tweets build a healthy relationship between police and citizens.

Report:

- These type of tweets must be responded urgently.
- Any person can be in danger, and had no way or time to go to police station. Therefore, police should make sure, no such tweet should be left unattended and response time should also be less.
- These type of tweets may not have imperative tone, may have words like “Urgent”, “danger”, ”report” etc. This information can used for modelling.
- Subjectivity and polarity of text can be used for modelling them too.
- Response times should be as minimum as possible.

Appreciation:

- These types of tweets help in motivating police staff.
- It also increases confidence of citizens on police department.
- Sentimental analysis can be used for modelling.
- These types of tweets generally have more likes than other types of tweets.
- High Response times on such tweets don’t affect much.
- Positivity of the post can also be checked using textblob.
- Words like “thank you”, “good”,”great” etc can also be searched for classification into such types.

Accountability:

- These types of tweets shows crimes, misguidance and misuse of power by police.
- This class is really important as it acts as a watchdog and questions actions of police.
- Sentimental analysis can be used to see how negative are the tweets.

- These tweets should not be left unanswered.
- These tweets make society more vigilant.
- These tweets also determine the confidence of society on the police department.

Other:

- These type of tweets include all other types which cannot be categorized into above classes.
- They can include important guildelines, spreading awareness on topics etc.
- They can include links, media for better interaction with the citizens.

b) How this classification can help police department.

- This classification will help police in analyzing tweets better. This will yield a better response time, more percentage of tweets will be attended.
- Police can directly iterate over all the reports as soon as possible as this class needs quickest attention.
- This will reduce their response time. As a result, people will get more connected to police as they are getting quicker responses from them.
- Many times people cannot go to police station. Specially, in today's scenario of covid, people avoid going to police station. So, when police is more responsive on twitter, it will bridge gap between them.
- Police can be held accountable for their actions; they can be cross questioned in front of all other citizens. This prevents a safer environment for citizens from some people of police department who try to misuse their power.
- Police can easily view tweets which are appreciating them, and can retweet them for a better reach. This will increase confidence of people on police.
- This increases reach of police among people, they can spread important guidelines in a better way.

At first I searched twitter manually, and found #COVIDSecondWaveInIndia trending on twitter.
So I collected 2000 tweets of this hashtag (tweets stored in "covidsecondwaveinindia.pkl").
Then, I found the top 10 hashtags mentioned in these tweets and found 2000 tweets for each of these ha shtags.

Code:

```
In [52]: def get_tweets_hashtag(hashtag,no_of_tweets):
             h=[]
             for s in tweepy.Cursor(api.search,q=hashtag,tweet_mode='extended').items(no_of_tweets):
                 h.append(s)
             return(h)
```

In [55]:
```
#finding top 10 hashtags
hashtag_count={}

for i in tweets_hash1:
    k=i.entities['hashtags']
    for j in k:
        j2=j['text']
        if(j2 in hashtag_count):
            hashtag_count[j2]+=1
        else:
            hashtag_count[j2]=1
top_10_hashtags=[]
for i in hashtag_count.keys():
    top_10_hashtags.append([hashtag_count[i],i])
top_10_hashtags=sorted(top_10_hashtags,reverse=True)
top_10_hashtags=top_10_hashtags[:11]
top_10_hashtags
```

In [ ]:
```
#collecting tweets for top 10 hashtags

section_2=[]

#automatic way
for i in top_10_hashtags.values():
    section_2.append(get_tweets_hashtag("#"+i,2000))

#manual way
tweets_1=get_tweets_hashtag("#COVIDSecondWaveInIndia",2000)
tweets_2=get_tweets_hashtag("#COVIDEmergency",2000)
tweets_3=get_tweets_hashtag("#COVID19India",2000)
tweets_4=get_tweets_hashtag("#COVID—19",2000)
tweets_5=get_tweets_hashtag("#SOS",2000)
tweets_6=get_tweets_hashtag("#ModiMadeDisaster",2000)
tweets_7=get_tweets_hashtag("#COVID19",2000)
tweets_8=get_tweets_hashtag("#bjphoardingremdesivir",2000)
tweets_9=get_tweets_hashtag("#BjpResponsibleForCoronaInIndia",2000)
tweets_10=get_tweets_hashtag("#CoronaSecondWave",2000)
tweets_11=get_tweets_hashtag("#COVIDSecondWave",2000)
section_2.append(tweets_2)
section_2.append(tweets_3)
```

Data collected:

Pkl files : section2_data.pkl, covidsecondwaveinindia.pkl

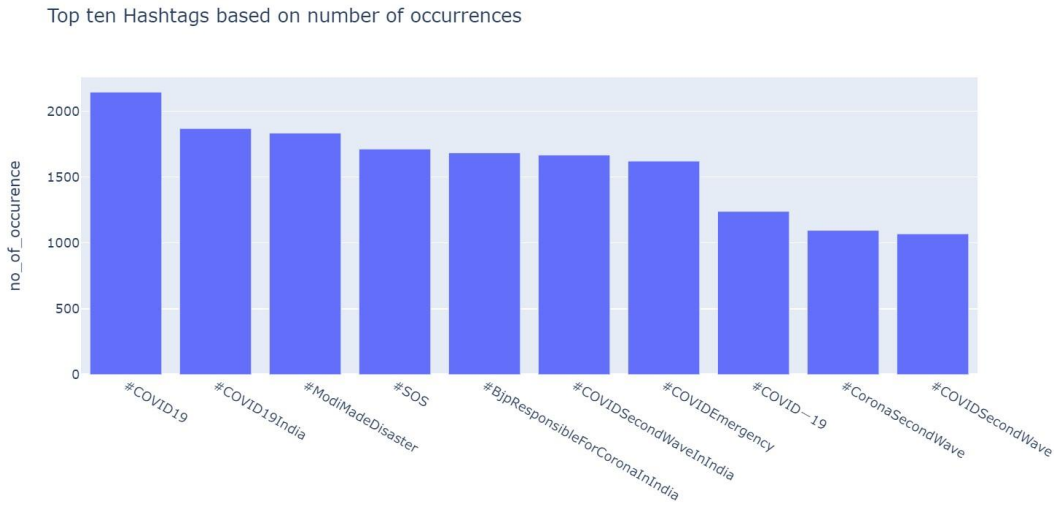
Top 10 Hashtags (exclude top hashtag)
[[1213, 'COVIDSecondWaveInIndia'],
 [333, 'COVIDEmergency'],
 [263, 'COVID19India'],
 [159, 'COVID—19'],
 [145, 'SOS'],
 [134, 'ModiMadeDisaster'],
 [99, 'COVID19'],
 [87, 'bjphoardingremdesivir'],
 [87, 'BjpResponsibleForCoronaInIndia'],
 [50, 'CoronaSecondWave'],

[46, 'COVIDSecondWave']]

Then I combined all 22000 tweets and removed duplicate tweets.

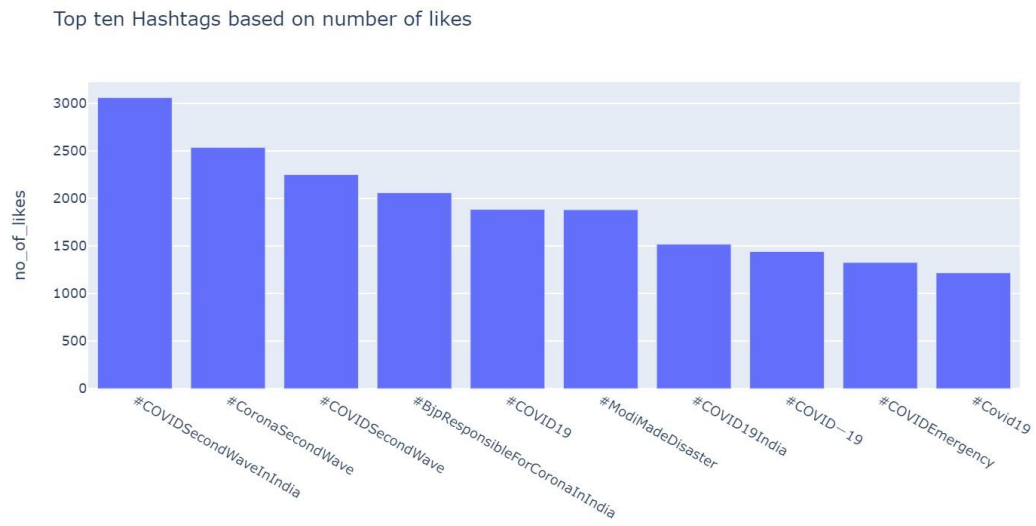
Top 10 hashtags based on most number of occurrences

	HashTags	no_of_occurence
12	#COVID19	2145
14	#COVID19India	1869
3	#ModiMadeDisaster	1834
26	#SOS	1713
2	#BjpResponsibleForCoronaInIndia	1684
0	#COVIDSecondWaveInIndia	1667
1	#COVIDEmergency	1621
5	#COVID—19	1239
33	#CoronaSecondWave	1095
79	#COVIDSecondWave	1069



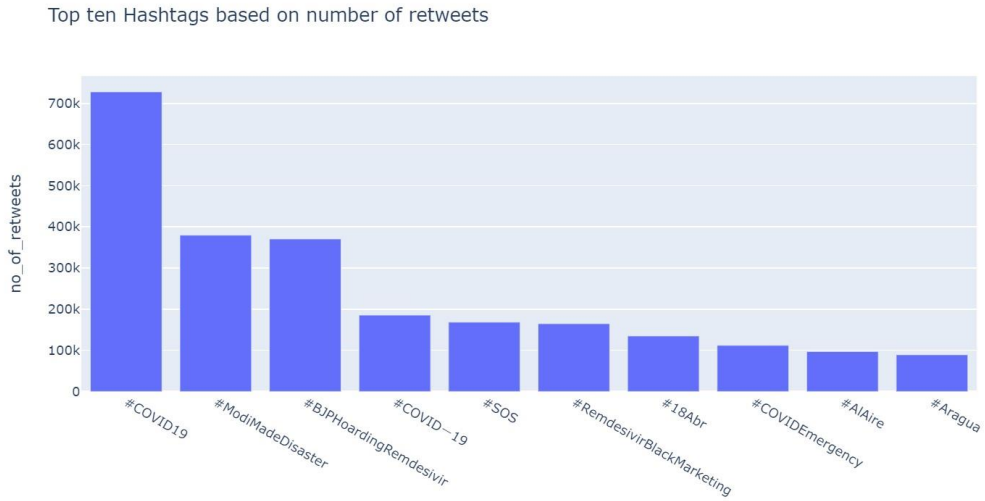
Top 10 hashtags based on most number of likes

HashTags	no_of_likes
#COVIDSecondWaveInIndia	3062
#CoronaSecondWave	2538
#COVIDSecondWave	2252
#BjpResponsibleForCoronaInIndia	2061
#COVID19	1885
#ModiMadeDisaster	1882
#COVID19India	1519
#COVID—19	1442
#COVIDEmergency	1327
#Covid19	1218



Top 10 hashtags based on most number of retweets

HashTags	no_of_retweets
#COVID19	728024
#ModiMadeDisaster	379761
#BJPHoardingRemdesivir	370570
#COVID—19	185538
#SOS	168370
#RemdesivirBlackMarketing	164661
#18Abr	134786
#COVIDEmergency	111982
#AlAire	97063
#Aragua	89111



Coefficient of Traffic Manipulation:

Set1:

{'#COVID19': 9.74308326926128,
'#COVID19India': 9.750312491765227,
'#ModiMadeDisaster': 10.097826668718346,
'#SOS': 13.313637337707922,
'#BjpResponsibleForCoronaInIndia': 8.326329618857669,

'#COVIDSecondWaveInIndia': 8.855381544236227,
'#COVIDEmergency': 10.975278296979017,
'#COVID—19': 12.331208950692702,
'#CoronaSecondWave': 10.968298408254412,
'#COVIDSecondWave': 9.661453936253341}

Set 2:

{'#COVIDSecondWaveInIndia': 8.855381544236227,
'#CoronaSecondWave': 10.968298408254412,
'#COVIDSecondWave': 9.661453936253341,
'#BjpResponsibleForCoronaInIndia': 8.326329618857669,
'#COVID19': 9.74308326926128,
'#ModiMadeDisaster': 10.097826668718346,
'#COVID19India': 9.750312491765227,
'#COVID—19': 12.331208950692702,
'#COVIDEmergency': 10.975278296979017,
'#Covid19': 28.200880606184718}

Set3:

{'#COVID19': 9.74308326926128,
'#ModiMadeDisaster': 10.097826668718346,
'#BJPHoardingRemdesivir': 19.114172022105134,
'#COVID—19': 12.331208950692702,
'#SOS': 13.313637337707922,
'#RemdesivirBlackMarketing': 33.16809116809117,
'#18Abr': 43.77261721329518,
'#COVIDEmergency': 10.975278296979017,
'#AlAire': 83.46376811594203,
'#Aragua': 24.871237603917947}

A table consisting of all 3 above sets with CTM values

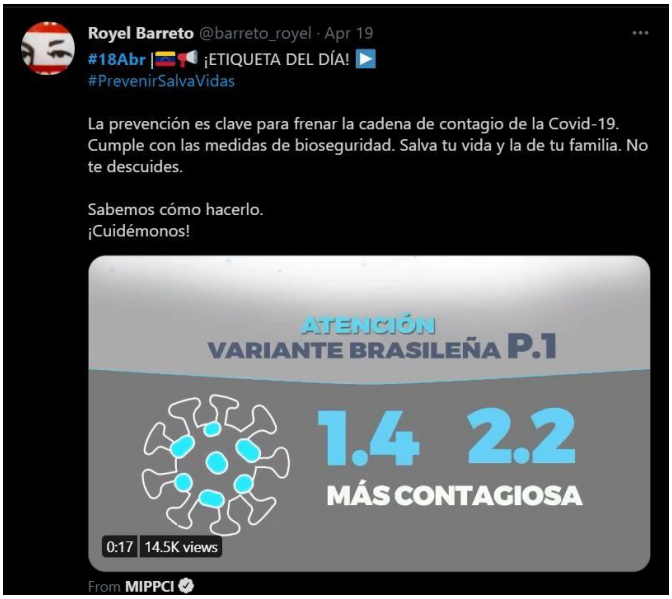
	Top_10_hashtags_occurence	Coefficient_of_traffic_manipulation1	Top_10_hashtags_likes	Coefficient_of_traffic_manipulation2	Top_10_hashtags_retweets	Coefficient_of_traffic_manipulation3
0	#COVID19	9.743083	#COVIDSecondWaveInIndia	8.855382	#COVID19	9.743083
1	#COVID19India	9.750312	#CoronaSecondWave	10.968298	#ModiMadeDisaster	10.097827
2	#ModiMadeDisaster	10.097827	#COVIDSecondWave	9.661454	#BJPHoardingRemdesivir	19.114172
3	#SOS	13.313637	#BjpResponsibleForCoronaInIndia	8.326330	#COVID—19	12.331209
4	#BjpResponsibleForCoronaInIndia	8.326330	#COVID19	9.743083	#SOS	13.313637
5	#COVIDSecondWaveInIndia	8.855382	#ModiMadeDisaster	10.097827	#RemdesivirBlackMarketing	33.168091
6	#COVIDEmergency	10.975278	#COVID19India	9.750312	#18Abr	43.772617
7	#COVID—19	12.331209	#COVID—19	12.331209	#COVIDEmergency	10.975278
8	#CoronaSecondWave	10.968298	#COVIDEmergency	10.975278	#AlAire	83.463768
9	#COVIDSecondWave	9.661454	#Covid19	28.200881	#Aragua	24.871238

Data inorganic or not:

- Since the CTM values for all the hashtags in the above table are lying in a moderate range, nothing shows that they have been generated inorganically.
- Since CTM is less, R/10 value is less, thus number of retweetes is not very high.
- Since CTM is less ,top 50 users are also not creating too much data, indicating that data is not inorganic.
- Since CTM is less , average number of posts by users is also less.
- Some hashtags like #ALAire has ctm of nearly 80. This shows that its data can be inorganic.
- Some hashtags like #18Abr has ctm of nearly 40. This shows that its data can be inorganic. However, there are few chances.

Inferences:

- Almost all the hashtags are related to COVID second wave in india.
- #ModiMadeDisaster, #BJPHoardingRemdesivir, #BjpResponsibleForCoronaInIndia are also related to covid issue as people are considering BJP and Prime minister Modi ji responsible for the second wave in india.
- #COVID—19, #CoronaSecondWave, #COVIDEmergency etc all are related to the same thing.
- #18Abr is somehow related to covid too (see a tweet below). Although nothing more can be said as it is present in some different language mostly by Columbian people.

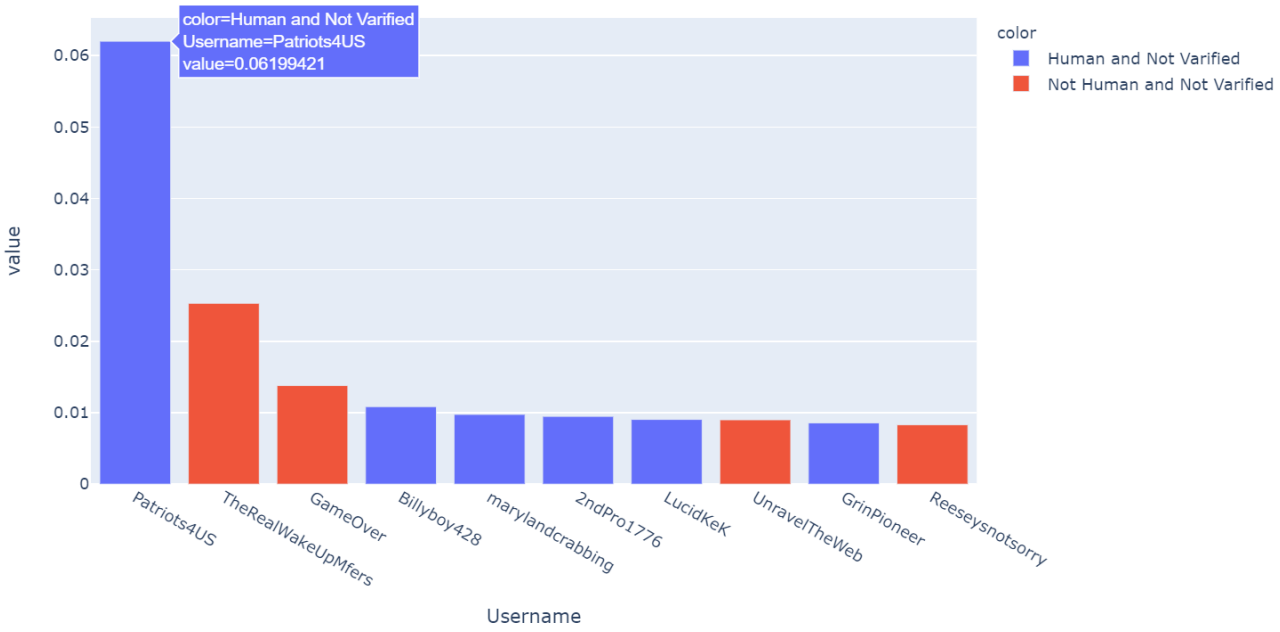


- Few unrelated hashtags can also be seen. #Aragua is not related at all it is related to some football league. Maybe covid has led to some changes in the league, that's why its also present here.

Parler

Collecting the top 10 usernames responsible for generating the most content.

a) A bar plot denoting their individual percentages.



Top Ten Users

```
Username
Patriots4US            0.061994
TheRealWakeUpMfers     0.025308
GameOver               0.013792
Billyboy428            0.010839
marylandcrabbing       0.009750
2ndPro1776             0.009476
LucidKeK               0.009044
UnravelTheWeb          0.008995
GrinPioneer            0.008563
Reeseysnotsorry        0.008299
Name: Name, dtype: float64
```

Top Ten Users Cumulatively

```
Username
Patriots4US            0.061994
TheRealWakeUpMfers     0.087302
GameOver               0.101094
Billyboy428            0.111933
marylandcrabbing       0.121683
2ndPro1776             0.131159
LucidKeK               0.140203
UnravelTheWeb          0.149198
GrinPioneer            0.157762
Reeseysnotsorry        0.166060
Name: Name, dtype: float64
```

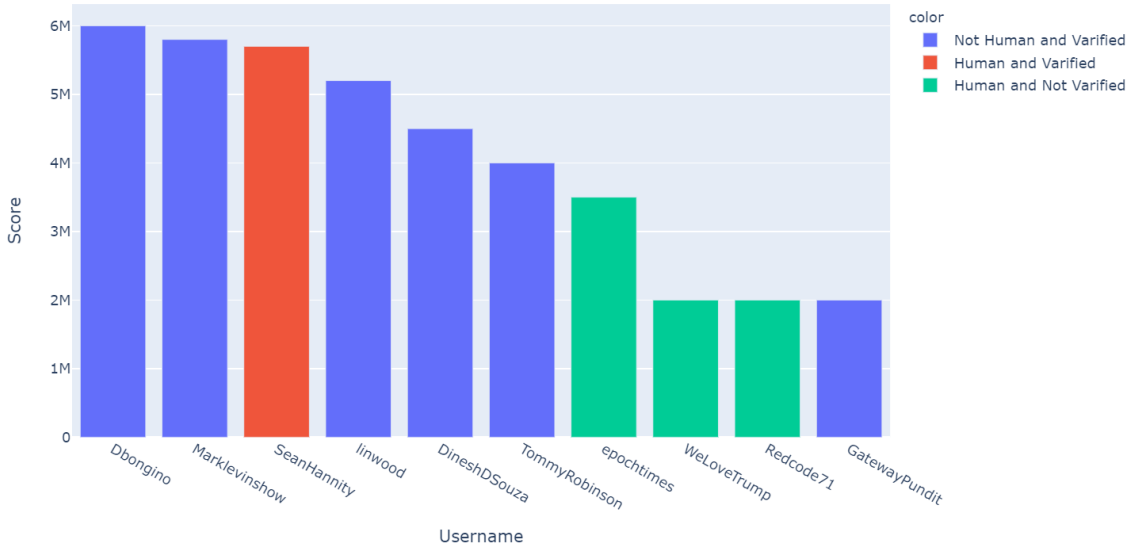
Cumulatively, 16.6060 % of the total content is generated by top 10 users.

Inference:

Patriots4US is a human user who is not even verified and he generates highest amount of co ntent that is 6.1994 percent of the total content generated. Four of these top ten content ge nerators are unverified bots.

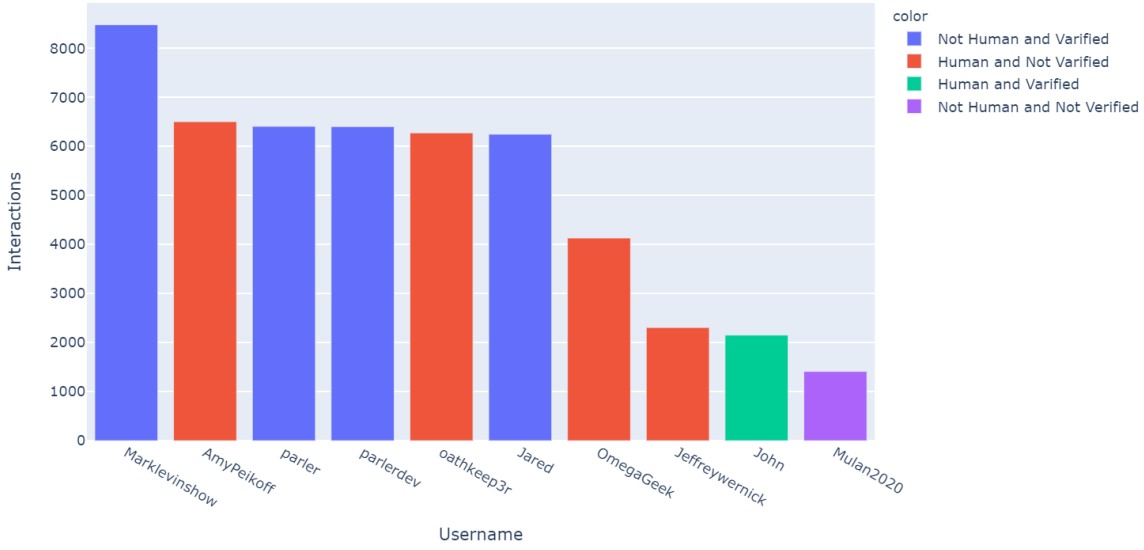
Out of the top ten users, four of them are bot accounts. None of them are verified accounts. This can be seen as an issue as out of these top ten users none is verified and so all the cont ent created by them has no verification/ credibility. This can increase fake data on parler if t here are no checks.

Top Ten usernames with most number of upvotes.



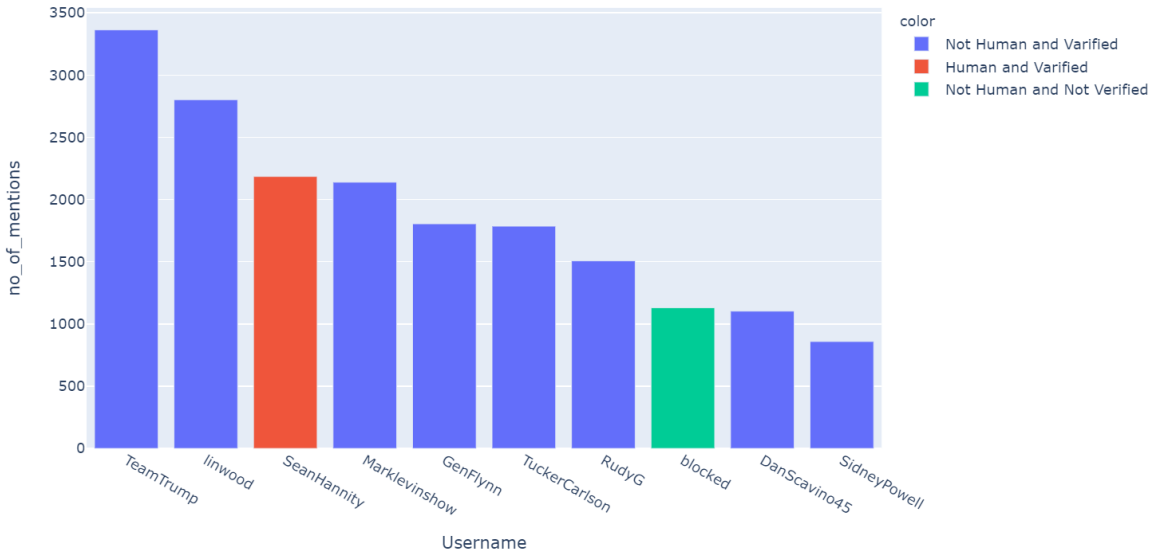
	Username	Score	Human	Verified
0	Dbongino	6000000	False	True
1	Marklevinshow	5800000	False	True
2	SeanHannity	5700000	True	True
3	linwood	5200000	False	True
4	DineshDSouza	4500000	False	True
5	TommyRobinson	4000000	False	True
6	epochtimes	3500000	True	False
7	WeLoveTrump	2000000	True	False
8	Redcode71	2000000	True	False
9	GatewayPundit	2000000	False	True

Top ten usernames with most number of interactions.



	Username	Interactions	Human	Verified
0	Marklevinshow	8478	False	True
1	AmyPeikoff	6503	True	False
2	parler	6407	False	True
3	parlerdev	6403	False	True
4	oathkeep3r	6275	True	False
5	Jared	6247	False	True
6	OmegaGeek	4131	True	False
7	Jeffreywernick	2307	True	False
8	John	2151	True	True
9	Mulan2020	1410	False	False

Top ten usernames with most number of mentions.



	Username	Human	Verified	Bio	no_of_mentions
19	TeamTrump	False	True	The official Parler account for the Trump Camp...	3363
43	linwood	False	True	NaN	2802
135	SeanHannity	True	True	TV Host Fox News Channel 9 PM EST. Nationally ...	2185
69	Marklevinshow	False	True	THIS IS THE OFFICIAL MARK LEVIN SHOW PARLER PA...	2139
52	GenFlynn	False	True	Digital Soldiers & Patriots Unite\n25th NSA to...	1805
...	...	...	...	...	...
8225	WWG1WGA4545	True	False	NaN	0
8224	Isaiah54vs17	False	False	Patriot\nConstitutionalist\nJesus is my King	0
8223	Kycsailor808	True	False	Member of a conservative minority living in Ka...	0
8222	JackMichaelJohnson	False	False	NaN	0
22325	Morloff175	False	False	NaN	0

22326 rows × 5 columns

a) A word cloud from the text in the bios of users.

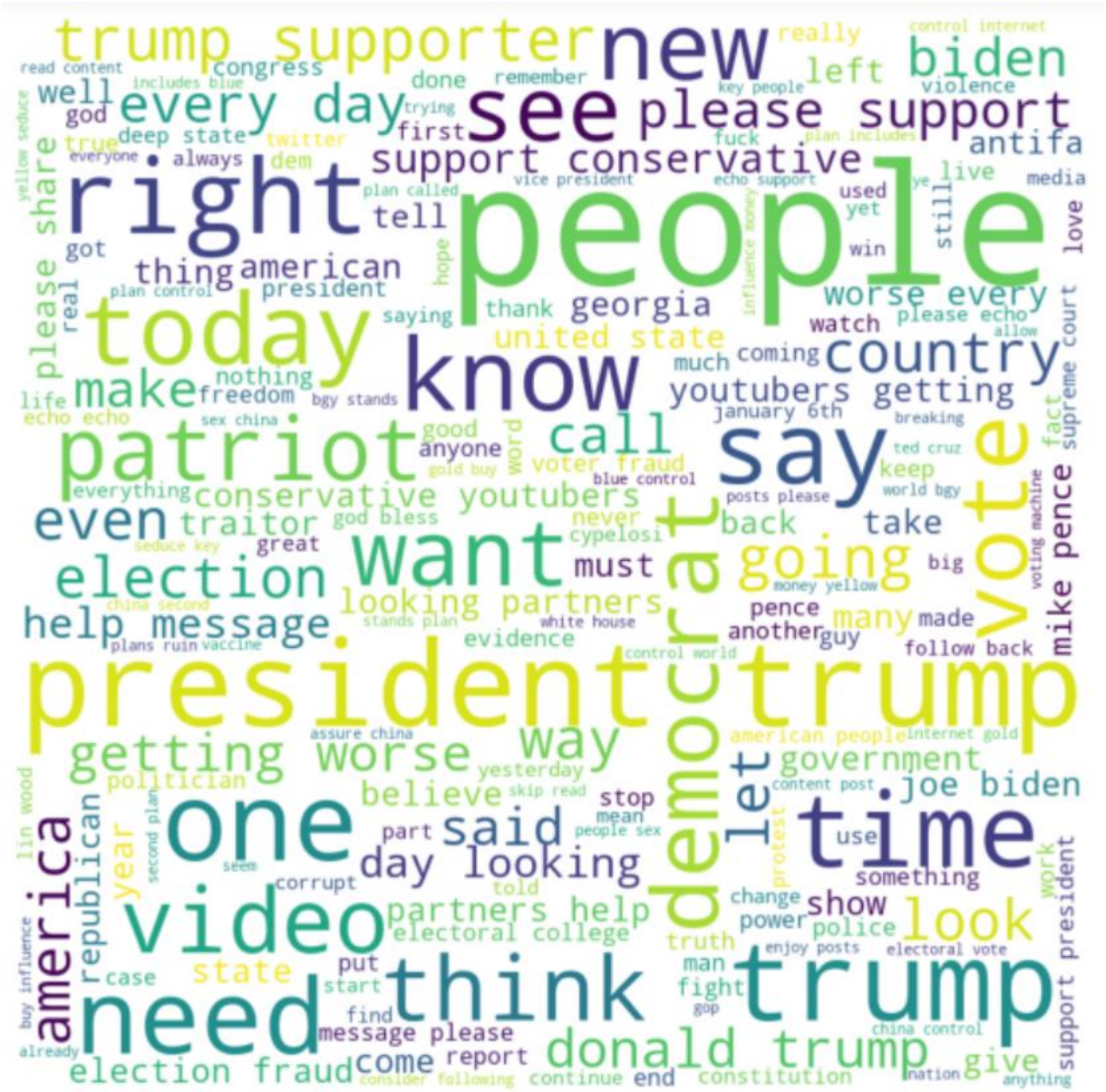
Inference:

Users with most number of interactions on parler can somehow be related to the parler company itself and maybe that's why ,the word "parler" is so frequent in

their Bio's. They may be holding some positions or have occupations which they might have discussed in their Bio's because of which we can see words like founder, investor, engineer etc. Since five out of these ten accounts are verified, their Bio's may have the word "official" to distinguish them from un verified accounts.

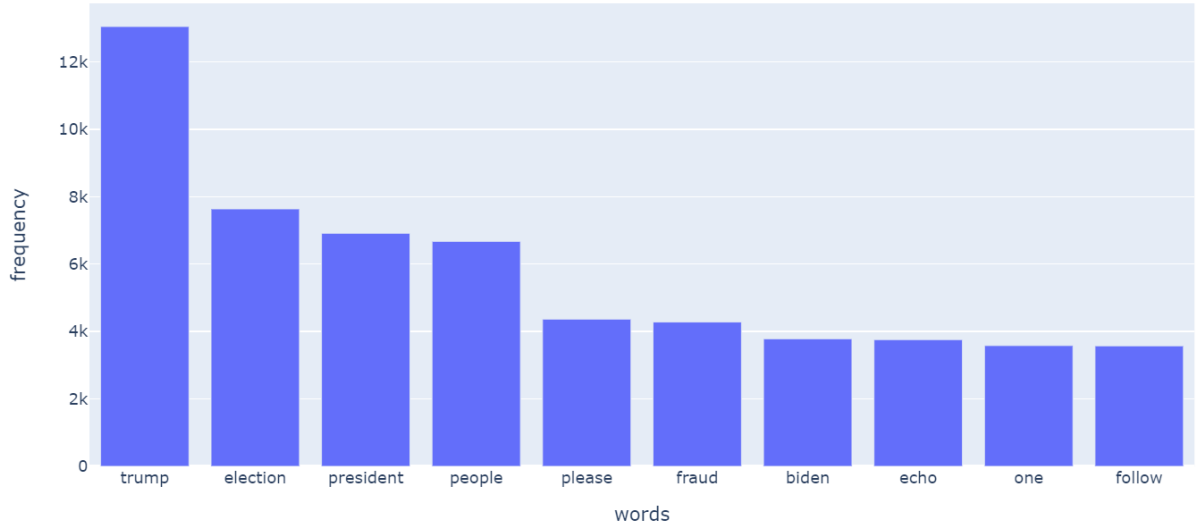
Identify the major topics being discussed in the content (posts) by extracting words and hashtags from the content.

a) Word cloud
Assumptions: for finding words, I am finding all alphanumeric words with length >2 and are not stop words.



Note: word “trump” has occurred highest number of times, still it is not the biggest in word cloud because of collocations. Therefore, instead of a single big word trump, we can see two words trump of decent size in our word cloud.

Top ten words based on number of occurrences



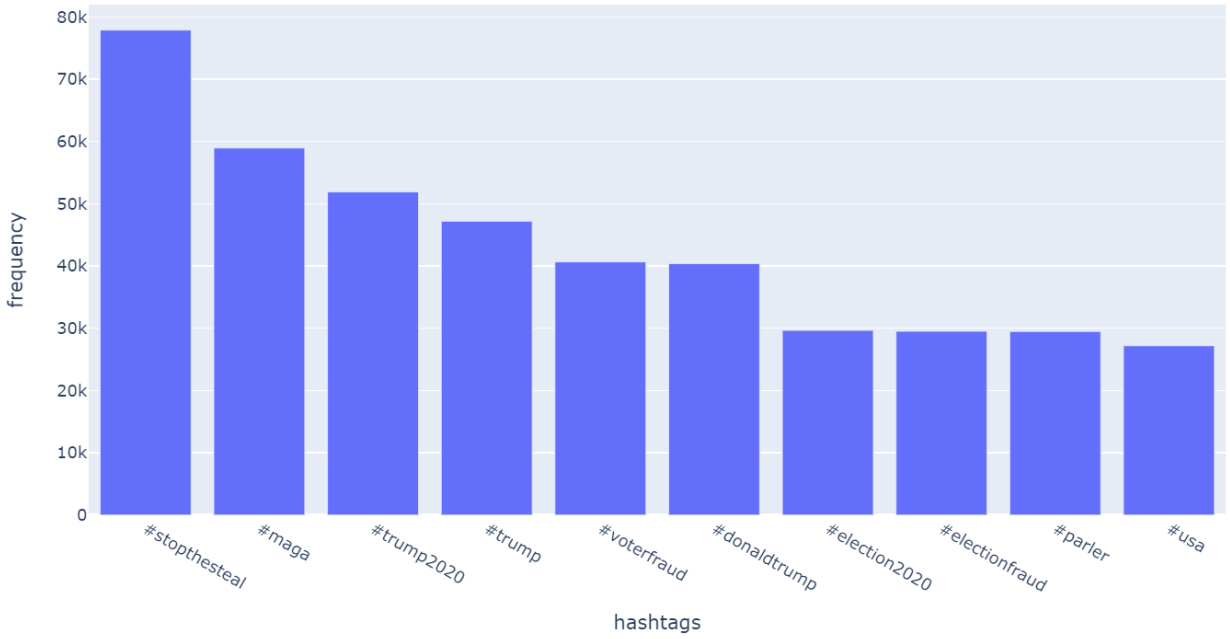
Out[11]:

	words	frequency
39	trump	13048
144	election	7637
246	president	6914
6	people	6674
938	please	4367
84	fraud	4280
72	biden	3783
137	echo	3758
288	one	3582
720	follow	3574

Observations/Inference:

Words like “trump”,” election”,” people”,” president”,” vote” etc shows that majority of posts are regarding voting in US presidential elections. Since word “trump” is bigger than “Biden”, we can infer that more posts are addressing Mr Trump. But we also see “Democrat” is bigger than “Republican” inferring more posts are related to democrats. By posts, people are supporting and criticising parties and people nominated for the post of US president.

b) Top 10 hashtags based on number of occurrences.

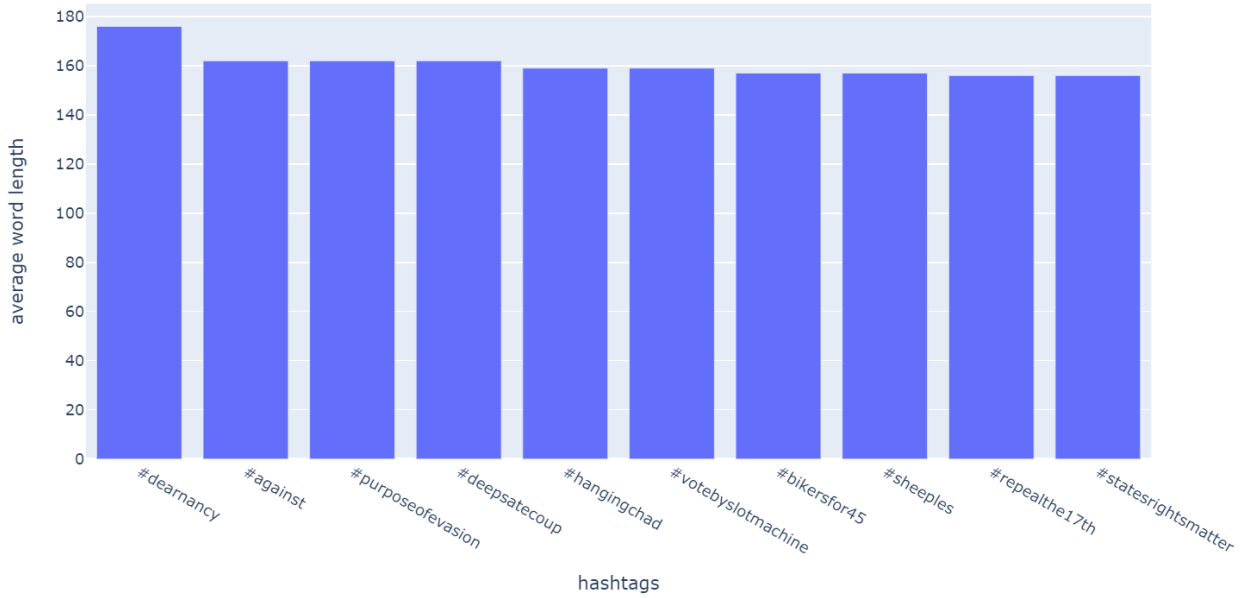


Out[16]:

	hashtags	frequency
15	#stopthesteal	77882
10	#maga	58939
55	#trump2020	51880
9	#trump	47204
12	#voterfraud	40644
56	#donaldtrump	40367
14	#election2020	29643
13	#electionfraud	29503
57	#parler	29473
111	#usa	27170

Top 10 hashtags based on average post length

Assumptions: For calculating post length, I am considering all the stop words and all alphanumeric words with length >1.



Out[17]:

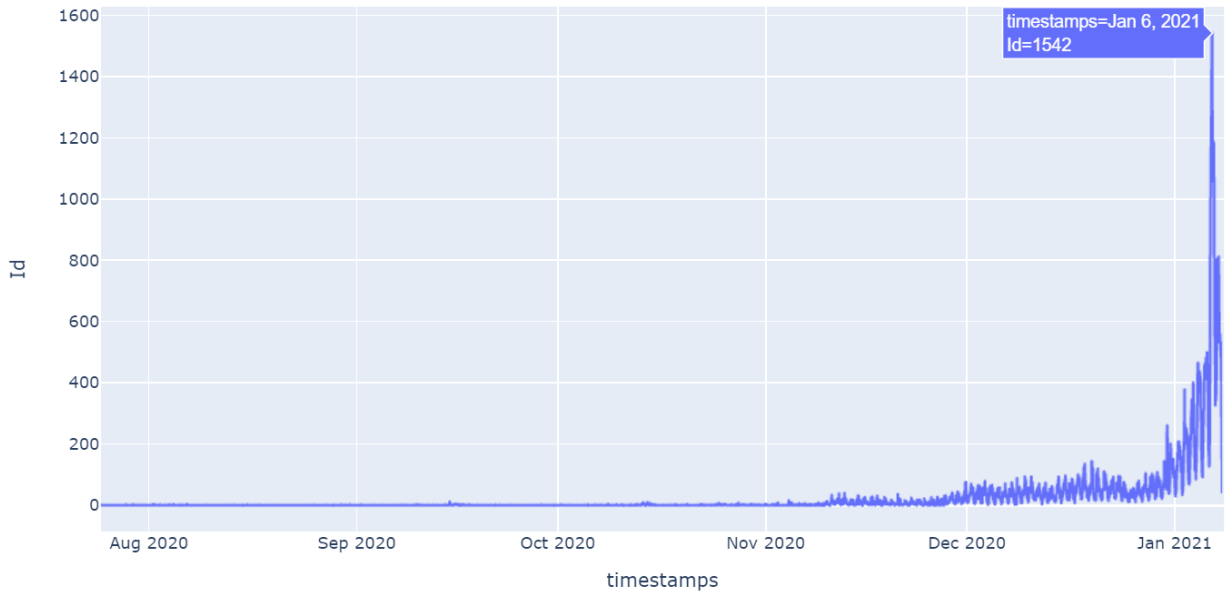
	hashtags	average word length
14612	#dearnancy	176.0
32101	#deepsatecoup	162.0
32102	#against	162.0
32103	#purposeofevasion	162.0
7025	#votebyslotmachine	159.0
7027	#hangingchad	159.0
25161	#sheeples	157.0
13016	#bikersfor45	157.0
26641	#statesrightsmatter	156.0
26642	#repealthe17th	156.0

Inferences:

In the top ten hashtags based on number of occurrences, almost all of these hashtags support Trump for US elections 2020. Some hashtags like #voterfraud convey the idea that people feel that there has been some voter fraud in elections.

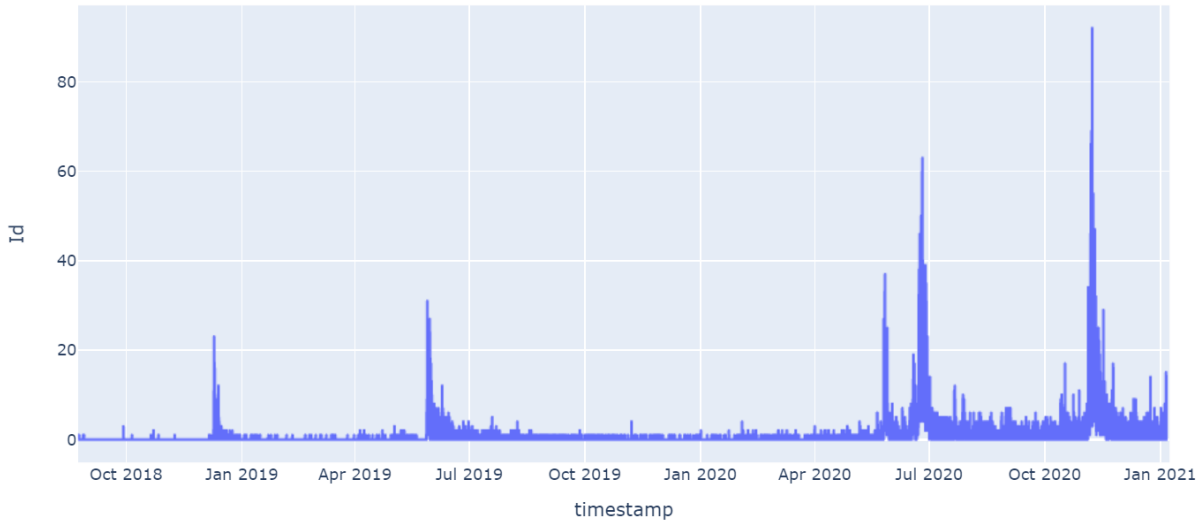
Trends in Parler Data

a) Content generation on parler



Assumptions: data have been clubber on hourly basis. Number of posts generated each hour can be seen on zooming the actual plot.

b) User account creation on parler



Inference (Content generation):

Before, Nov 2020 there was not much content generated. As elections approached, there was a gradual increase in number of posts. Then on Jan 6 2021, we see a peak of content generation on parler. On checking Internet, I found that the peak of content generation came on the same day which marks "The storming of the United States Capitol".

Inference (Account creation):

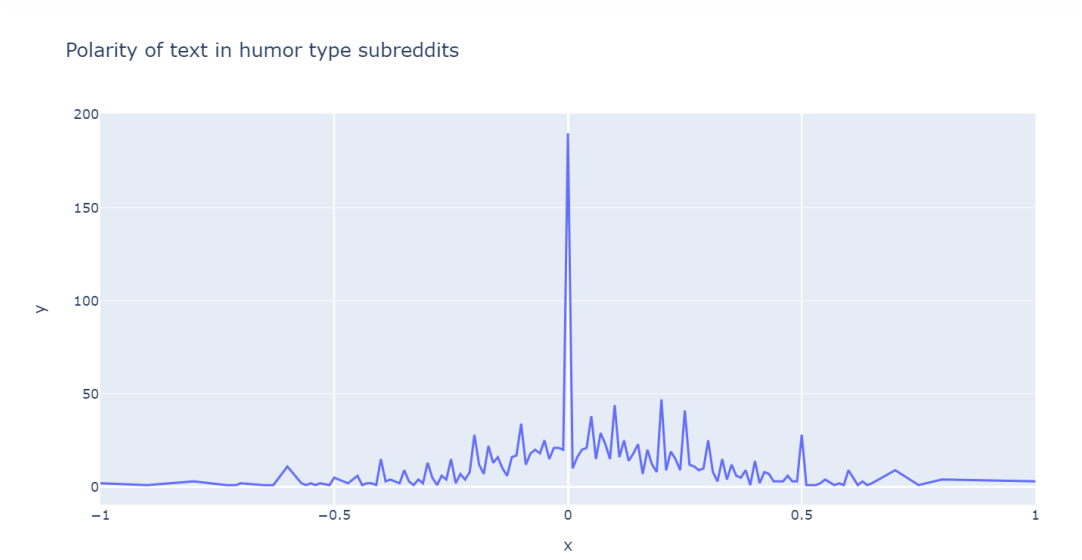
It follows the same pattern as graph of content generation, but has many local peaks. The biggest peak comes at the time of us presidential elections. More accounts are getting created in the late 2020 and Jan 2021 than before.

Reddit

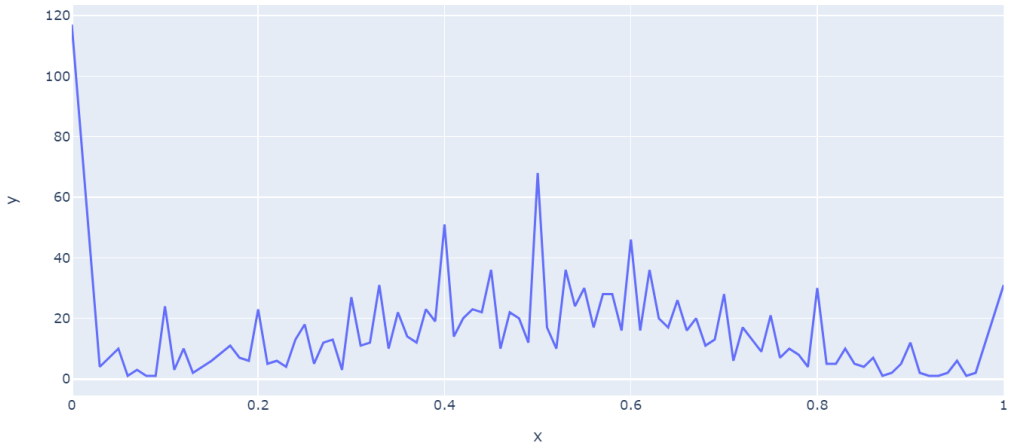
Section I

Across the two categories of subreddits, preprocess the text corpus and analyse it using the following methods:

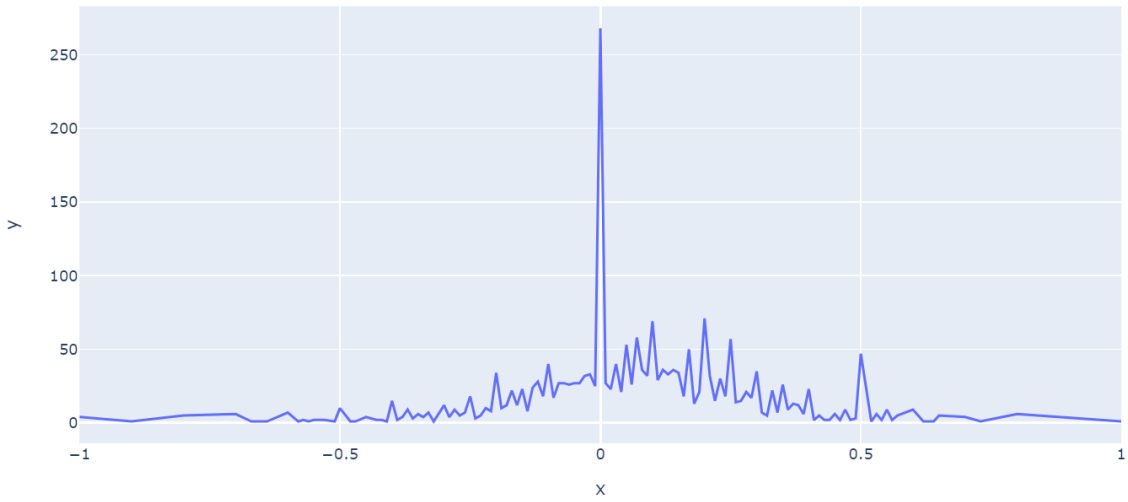
a)

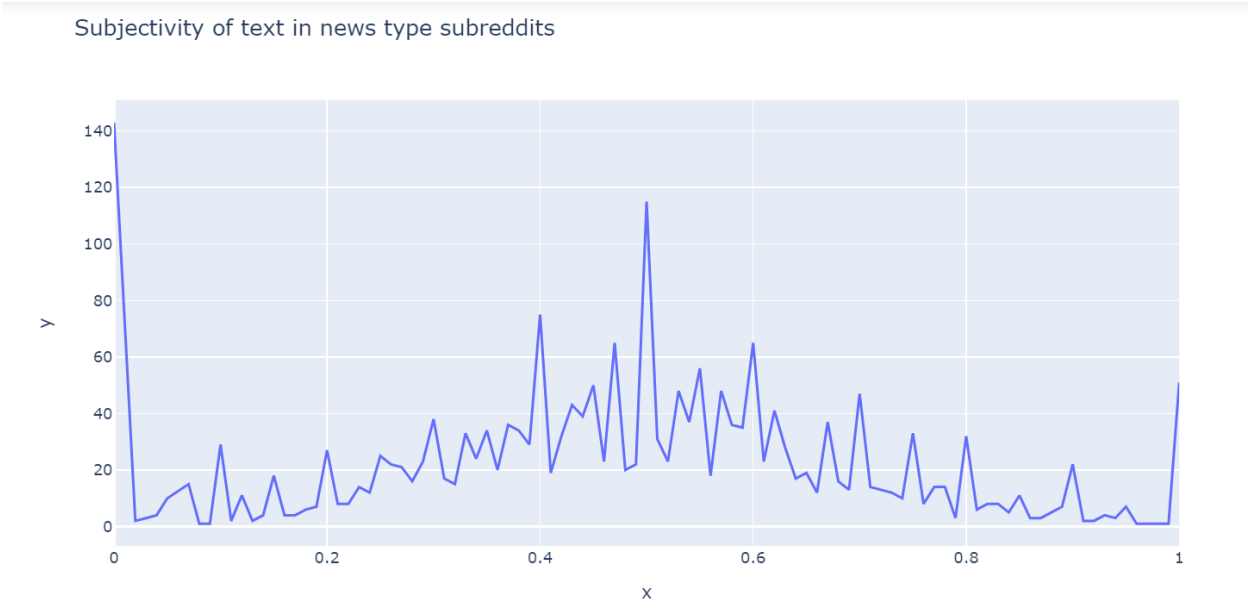


Subjectivity of text in humor type subreddits



Polarity of text in news type subreddits






```
Humor:
Polarity
Mean:  0.05158728575619289
Standard Deviation:  0.24874836482781126
Subjectivity
Mean:  0.47140332998753287
Standard Deviation:  0.24277567078204632
------------------------------
News:
Polarity
Mean:  0.0640485348011969
Standard Deviation:  0.2304684802126608
Subjectivity
Mean:  0.4704749753548108
Standard Deviation:  0.2327991434383808
------------------------------
```

Observation:

- For text in humor type subreddits graph, we see that there is a huge spike at polarity=0. Ie good number of data don't show any inclination to either 1 or -1. These data(polarity=0) are showing a neutral sentiment. We can also see that majority of spikes are near zero(x=0.05,0.1,0.2-0.1 etc). this shows data has a neutral sentiment. We can also see more positive sentiment peaks than the negative one. Negative peaks can represent dark humor, controvertial humor too.
- For text in humor type subreddits graph, we see in subjectivity, huge peak at subjectivity=0 and significantly big peak at subjectivity=0.5. for subjectivity=0, we can say that a lot of data is totally objective(clear cut). We see peak at x=0.5 and many smaller peaks near in the middle of x axis. This tells that good amount of data is fairly subjective.
- For text in news type subreddits graph, we see polarity peak at x=0 is bigger than that we got in polarity graph of humor subreddit. This shows that data in news has

more neutral sentiment. We can also see more positive sentiment peaks than the negative one.

- For text in news type subreddits graph, we see that more data zero subjectivity in news subreddit than that in humor subreddit. Thus news data is more objective. Also, the peaks in middle of x axis(x=0.4,0.47,0.55,0.6) are generally bigger than that in humor type subreddit subjectivity graph.

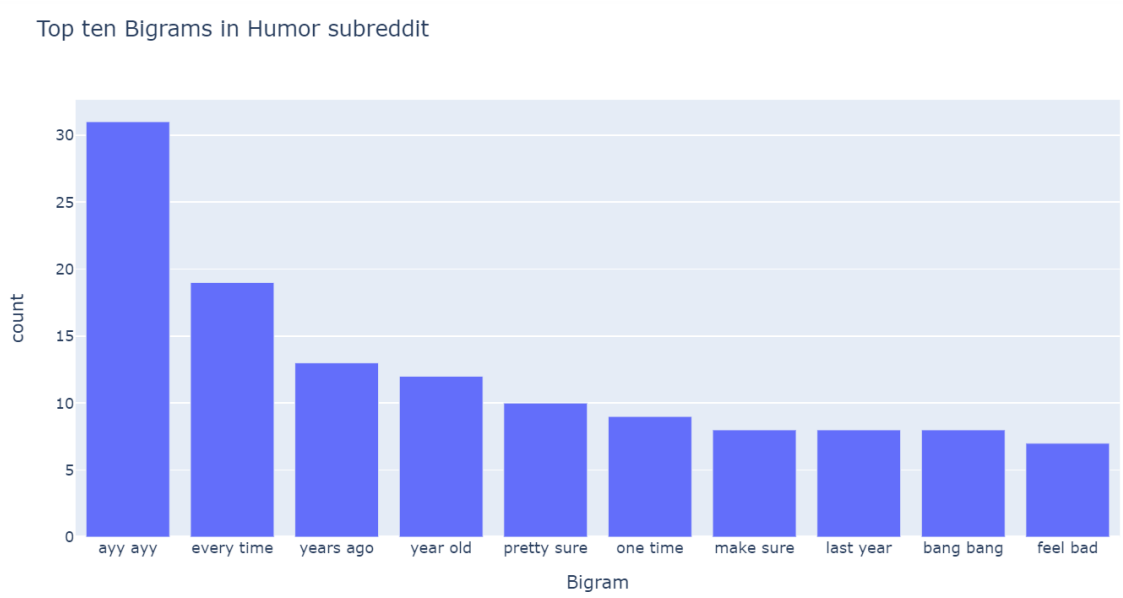
Median/standard deviation observation:

- Median for humor type subreddit data is 0.05(approx) while for news type subreddit is 0.06(approx)
- there is not much significant difference but one thing can be said is that data in both types shows majorly neutral sentiment.
- However, news data is very little to the positive side than humor type data.
- Standard deviation for humor data is a bit more, this means overall data deviates a bit more from mean in humor type data than news type. However, it is very minor difference.

Plot distributions of the top 10 most occurring (i) unigrams, (ii) bi-grams across the two categories.

Assumptions:

- Any word present in the stop words set or had length<=2 was considered "unnecessary" word by me.
- Any bigram/unigram having such a word was excluded from top 10.

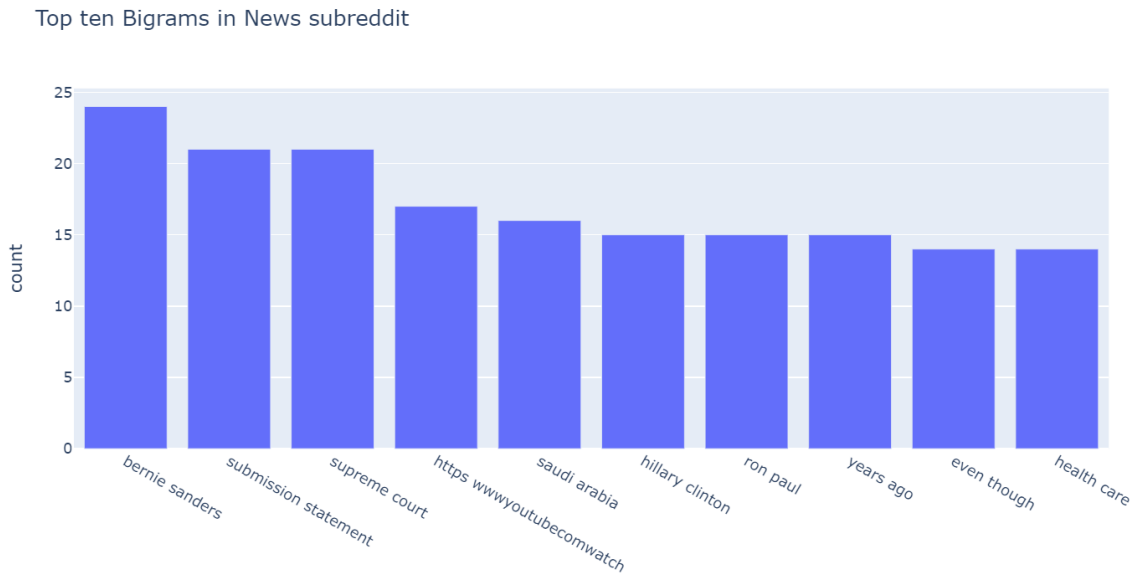


Inferences:

- We can see that word ayy ayy tops the graph.
- Since its data of humor subreddit, ayy ayy can be a way to greet others in some way which people are majorly using
- . Words like years ago, one time , last year etc may have been used to point on some funny past experience.
- Word bang bang is also a type of slang which is maybe used by people in some funny way.

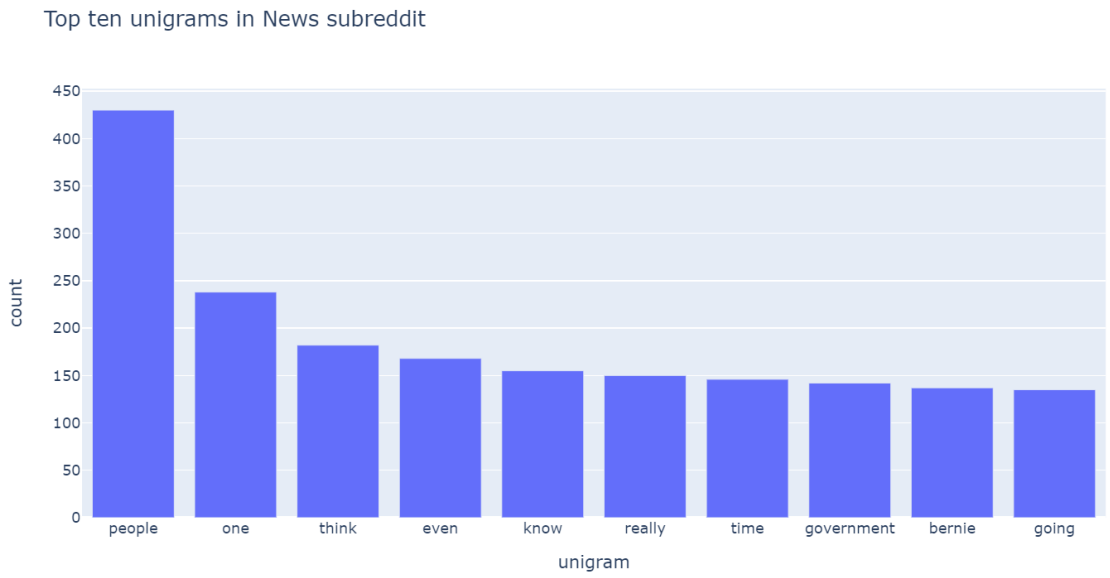
Inferences:

- Words like people, time, guy are common words people might be using for referring people, crowd in a post, stand-up or a joke.
- “Really” might be used to subjectify things which is needed in humour.
- got can be short form of game of thrones if its not a stop word, and we know there is so much humorous content out there forgot because of season 8. However, it’s difficult to make more inferences out of unigram data.
- Most of the words are common words.
- Bigram’s data provided more information.



Inferences:

- Bernie sanders is an American politician(senator) and a former activist who was a lot in american news because of covid measures and many other issues. That's why he tops the chart.
- Supreme court might have been in news as its one of the most important institution in any state.
- People might have shared youtube videos which is why youtube.com is also present in the graph.
- Hillary Clinton is also a well renounced politician and can be there in news.
- Saudi arabia (country with oil reserves)is so called us ally and chances are there it can be present in news.



Inferences:

- Just like top bigrams of news, bernie of bernie sanders is present there in the top 10.
- Government related, people related news comes in mainstream and so these words are present there.
- Other words are common words and can be present in news articles.
- Bigrams for news had more information.

Build a machine learning classifier to classify the comments into the two categories "news" and "humour".

a) Steps:
   - I used train_test_split function to split data into 70:30 ratio of train test
   - CountVectorizer was used to make bag of words out of training data.(We were allowed to use libraries)
   - Training data was converted into vectors using transform function. X_test was also transformed.
   - Y was changed. 1 for humor and 0 for news
   - Multinomial naïve bayes classifier of sklearn was used and trained on x_train, y_train.
   - Then model predicted y_predicted
   - Y_predicted,y_test was compared to find accuracy, precision, recall, f1 score and confusion matrix.
   - This model was used as it was giving good results in this type of data and worked fine with huge feature set. For text classification, multinomial Naïve Bayes is widely used.

b) Model named “"bag_of_words_model"” can be found in zip.

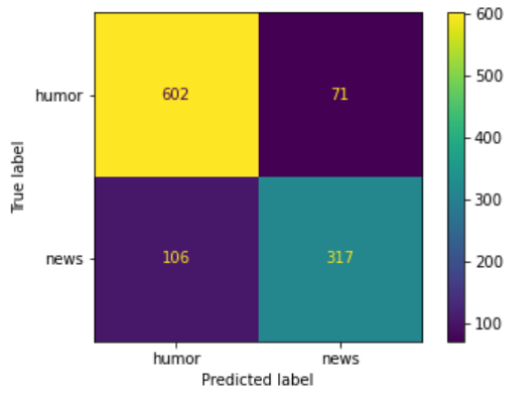
c)

```
Accuracy: 0.8385036496350365
Precision 0.7494089834515366
Recall 0.8170103092783505
F1 score 0.7817509247842169
```

We can see that accuracy is more than precision, recall and f1 score.

Confusion Matrix:

```
array([[602,  71],
       [106, 317]], dtype=int64)
```



d)

5 wrongly predicted text

```
as a bears fan  i have heard this as  if your parents were morons  then i d be a packer fan  1
----------------------------------------------------------------------
i lost interest after college dropout he seemed to lose the flair that made him unique and grew big headed ca nt stand the guy
now  what kind of asshole hijacks someone else s moment at the music awards  1
----------------------------------------------------------------------
when people use that term  i ll ping you  if i do nt punch them in the face  is that considered packet loss  1
----------------------------------------------------------------------
we always hear about these scientific breakthroughs on reddit  and personally  i never see them anywhere else then  we never he
ar about them again why  0
----------------------------------------------------------------------
the more i think about it the more i m disappointed that i did nt have the post say 8  46am 1
----------------------------------------------------------------------
```

Note:

“0” after text denotes text belonged to news subreddit but predicted as humor subreddit

“1” after text denotes text belonged to humor subreddit but predicted as news subreddit

5 correctly predicted text:

```
she ll be jailed only if there are other factions with enough sway that want to see her jailed it s looking like the intelligen
ce  faction wants that i m holding out hope jailing her wo nt really change much in and of itself  but it s a nice dream  0
-----------------------------------------------------------------------
i really do hate the warhawkishness of this sub ya ll are getting war  all in my liberty  and i hate it  0
-----------------------------------------------------------------------
of course it is at this point literally everything  and i do not mean  literally  like a valley girl uses it  i mean just consi
der everything to be racists from now on please  and stories like this will not come as a surprise 0
-----------------------------------------------------------------------
 100 for winning employee of the quarter  sounds pretty reasonable what am i supposed to be mad about here  frugal management
0
-----------------------------------------------------------------------
  the acting director  army public relations  col sani usman  said in a statement on monday that the troops of 26 task force ba
ttalion  and cameroonian forces  also rescued 112 captives of the boko haram during a joint operation at kirawa junction  and t
hat the rescued comprised eight men  36 children and 68 women    https  wwwtodayngnewsnational80255camerooniantroopsrescue112bo
koharamscaptives  0
-----------------------------------------------------------------------
```

0” after text denotes text belonged to news subreddit but predicted as news subreddit

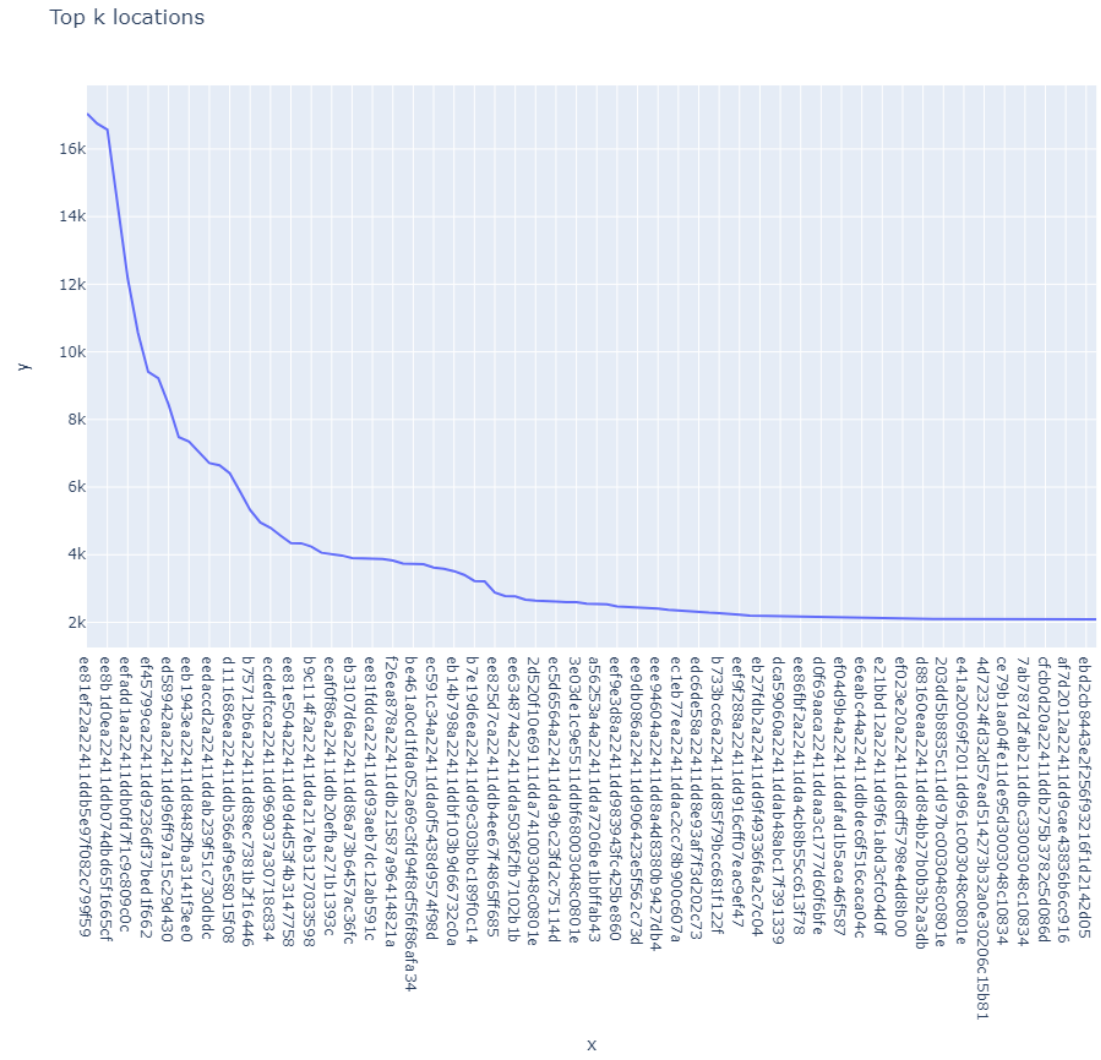
“1” after text denotes text belonged to humour subreddit but predicted as humour subreddit

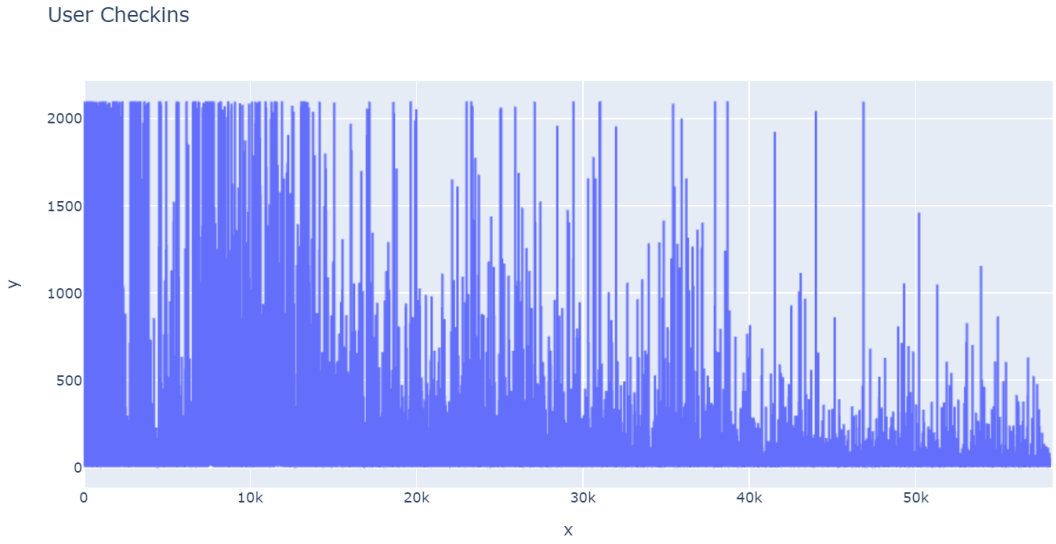
<u>Reasoning:</u>

- They were predicted wrong because they have many words which belonged to opposite subreddit.
- Our model predicts a sentence on the basis of words it has and the weight of these words.
- If a word belonging to news sub reddit comes in a sentence of humor subreddit, then that sentence can be classified as humor.
- Here also, sentences which were classified wrongly had words of opposite subreddit.
- Occurrence of some words (belonging to opposite subreddit) in high frequency also makes our prediction wrong.

BrightKite

a)



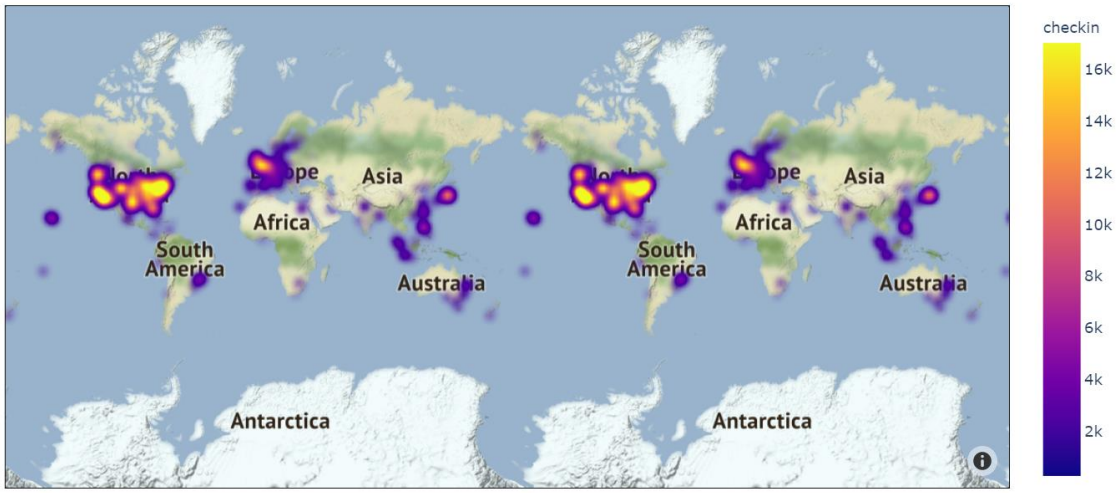


Note: I have printed distribution of all users (not top k users) as it is an interactive plot and can be visualised better by zooming and pointing.

Inference:

- In Location distribution graph, We can see that user with topmost location has been checked 17.049 k times.
- There are even locations which have been checked in just 2k times.
- There is a huge range in number of checked ins which shows some places are highly popular among users.
- User check in distribution graph, I have plotted users with number of check ins. There are many spikes in the graph showing that many users have high number of check in compared to others.
- There are many users with 2100 check-ins (in hundreds).
- Many user have check ins in just hundreds. This is a big range of data.
- Location ID 'ee81ef22a22411ddb5e97f082c799f59' has highest number of check-ins.

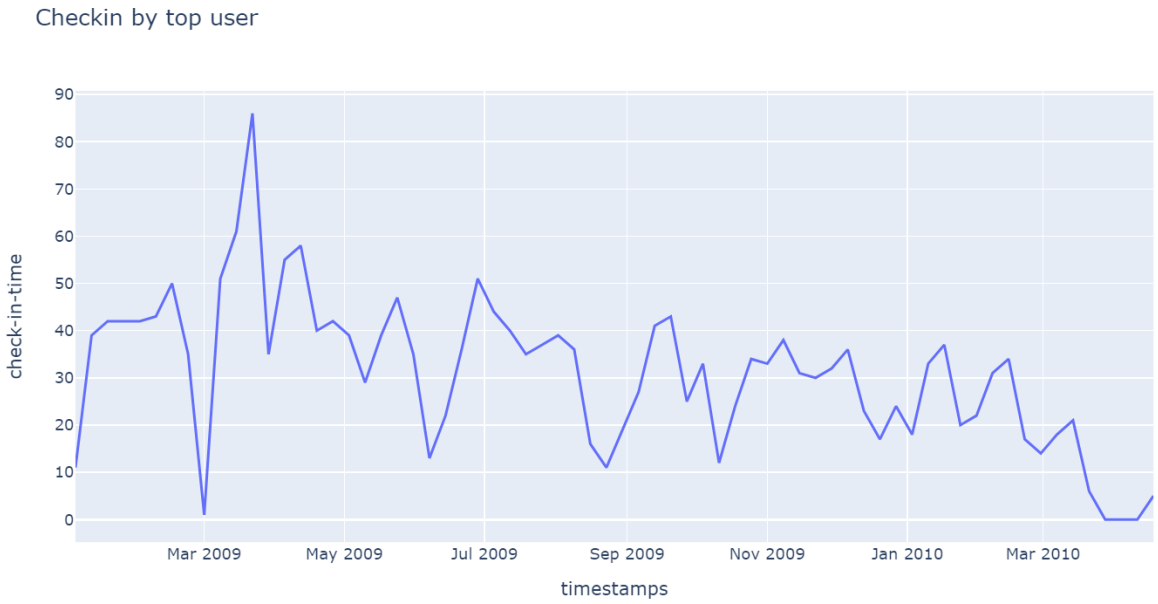
b) Top 2000 places with most check-ins

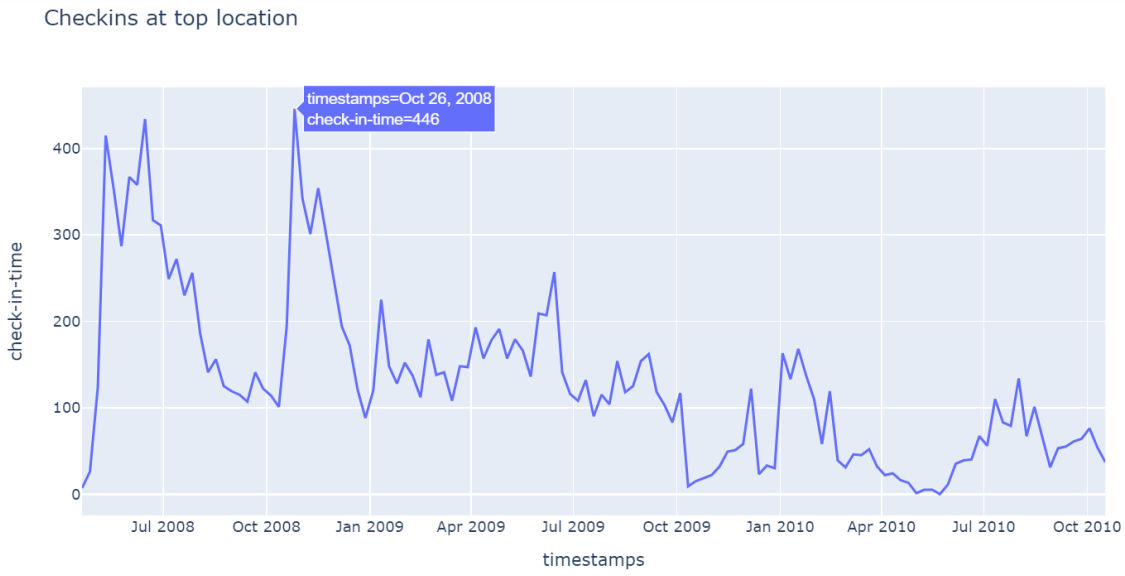


Observations:

- We observe major check in in Europe, north America, south east asia
- very small patches are visible in Africa, india, australia.
- A large part of asia don't have any patches.
- This shows that its not popular in asia except the south east part.
- We can see more patches near tropics than colder regions too. This shows that this app is less used in colder regions.
- For north American, this app is more used in countryside as its more dark there.
- Similarly for Europe, yellow part(less usage) is present in the middle of highly used areas(dark region).

c)





<u>Observation:</u>

Temporal information can be used to target users.

- For top user, one can see that he has few check in on march 1 and sudden peak on march 22. This info shows when is the user free and when not.
- This may also mean that user might have been busy from march 1 to march 22 in something like exams etc. To sum up one can extract info which user don't want others to know.
- For top location data, we can see that year-by-year users check ins on top location is declining.

- In June 15,2008 there were 434 check-ins and on June 14,2009 there were 257 check-ins. This data can be used with other data to get user reasons because of which location is becoming unpopular.
- Hotel, motel owners can see when users are most and then they can charge them high rates for staying.
- Other businesses can also rate users more based on popularity of location at that time of season.

Yes, the data is periodic. Peaks are present near may-June and sept- nov every year on top location check-in graph. For other part of year there is a downfall.

Leveraging information

- This info shows when is the user free and when not depending on the number of checkins. So, companies can design their policies and trading plans according to it to gain more profits.
- Brands will know when a user is busy and when he is ready to go to different places. They can advertise then trip plans etc on those particular parts of year.
- Brands can know even if you are into adventure, or like 5-star services depending on your check in locations and can sell u necessary equipment.
- Hotel, motel owners can see when users are most and then they can charge them high rates for staying.
- Other businesses can also rate users more based on popularity of location at that time of season.

Privacy issues:

- Just like we saw in previous part for top user, one can see that he has few check in on march 1 and sudden peak on march 22. This info shows when is the user free and when not. This may also mean that user might have been busy from march 1 to march 22 in something like exams etc. To sum up one can extract info which user don't want others to know. This hinders right to privacy.
- Data can be sold to other companies who can use it to lure customers on the basis of their behaviour.(eg showing equipments which they like)
- Data can be used to manipulate users too.
- Our location is one of our most valuable information. This can be used by criminals too. Even if company don't provide this data to criminals, they can take it out via data preaches.