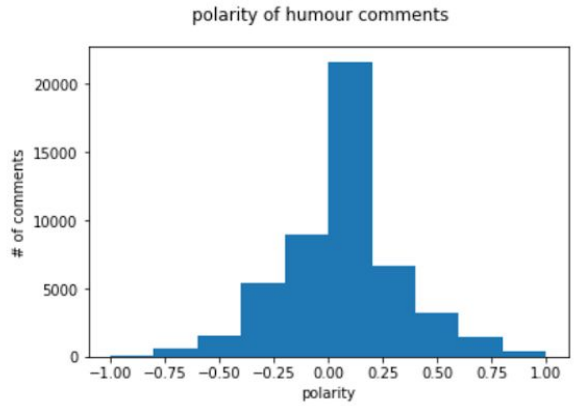
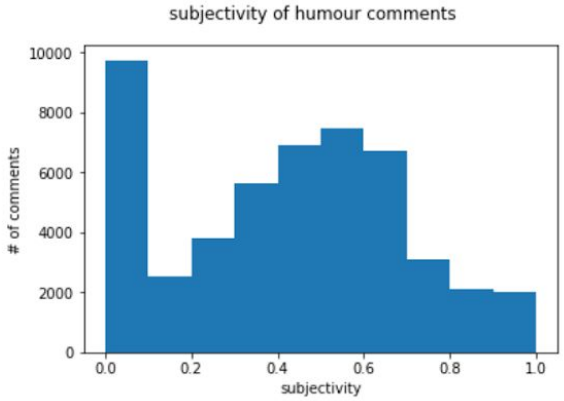
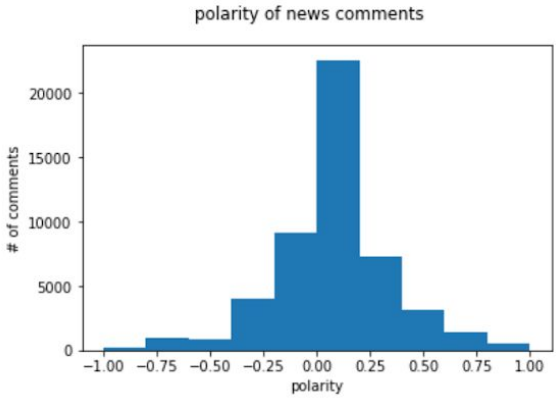
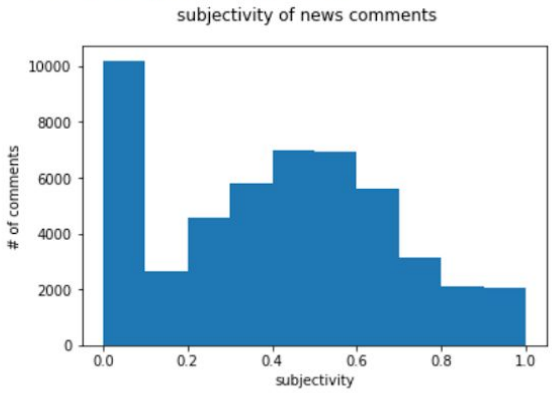
PSOSM MIDSEM
Rishabh Devgon | 2018303

Section - I
Q1. a)





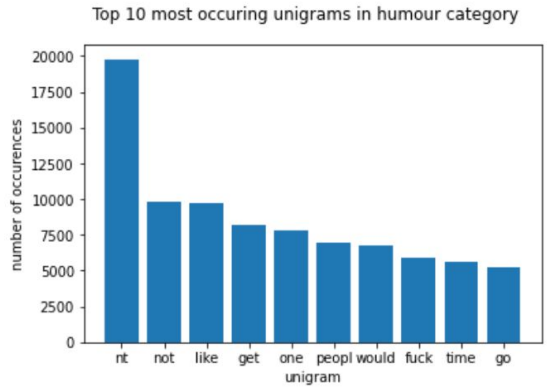


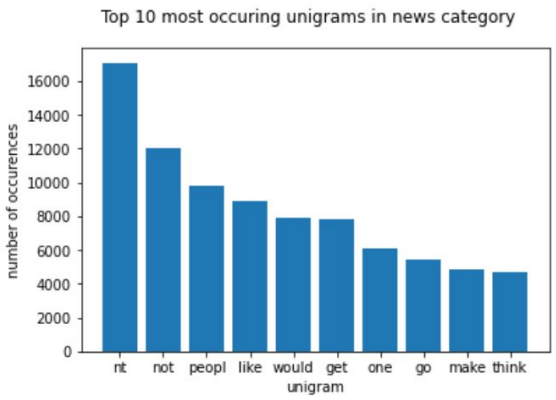


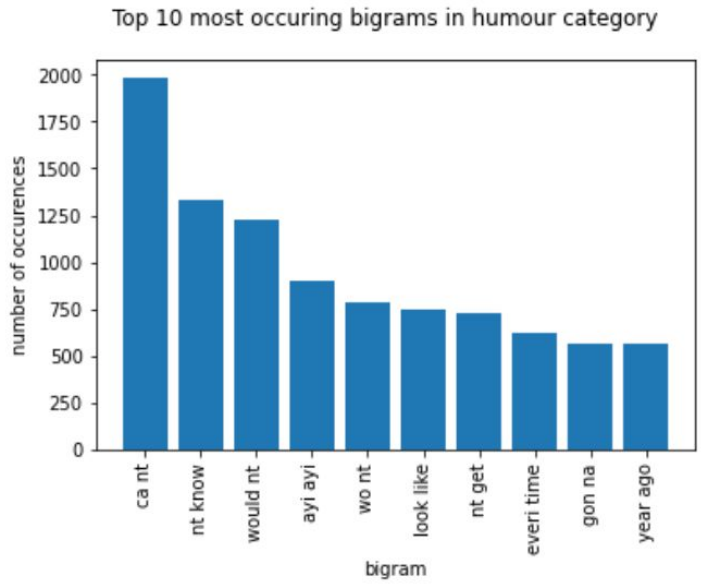

```
The mean of the polarity of humour comments is 0.03689885524022274
The standard deviation of the polarity of humour comments is 0.27707355891141133
The mean of the subjectivity of humour comments is 0.4097064284222127
The standard deviation of the subjectivity of humour comments is 0.27362188397090936
The mean of the polarity of news comments is 0.050583834095283026
The standard deviation of the polarity of news comments is 0.2733419067626938
The mean of the subjectivity of news comments is 0.39816051842180306
The standard deviation of the subjectivity of news comments is 0.2744939660121272
```

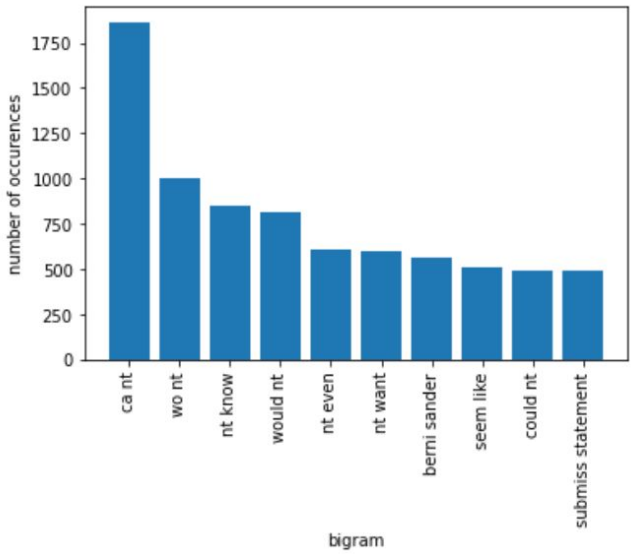
The polarity in reddit seems to follow somewhat of a gaussian curve with most comments lying between -0.25 and 0.25 polarity. However, subjectivity is more sporadic in the graphs with the most popular subjectivity being around 0 and 0.1. There is not a lot of difference between news and humour in terms of subjectivity and polarity because the graphs look almost identical and so do these first order statistics. This depicts a sentimental similarity between news and humour.

b)









The most occurring unigrams in humour category contain ‘people’ and ‘time’ which means that a lot of humour takes these subthemes and also contains the word ‘fuck’ which depicts the popularity of this word in the reddit vocabulary. In the news category, again ‘people’ seems to

pop up quite a lot which shows it's popularity in usage which makes sense given how news is reported. Besides this, coming up with observations from unigrams could be a bit of a stretch given its nature. The bi grams in the humour category show the usage of 'look like' and 'every time' which sort of acted as joke/ meme templates. There is also 'ayi ayi' which seems to have a lot of usage but on searching this term online, I could not quite understand why it was so popularly used. In the news category, we observe that submission statement and Bernie Sanders is widely used thus pointing to the political nature of the news comments on Reddit.

Q2. a. I have used alphabets used in the english language to analyse the text to not add to unrequired computational overhead. I have also converted all the text to lower case for the sake of uniformity and to not make words with case difference be treated as separate words. I have then removed all stop words in the english language to avoid any words that may not add any meaning to the text corpus but create computational overhead. I have also used lexicon normalisation to make the text more uniform for the model to analyse.
b.

```
from sklearn.linear_model import LogisticRegression

clf = LogisticRegression(random_state=0).fit(X_train, y_train)
y_pred_lr=clf.predict(X_test)
print(y_pred_lr)
```

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.30, random_state=42)
```

```
from sklearn.linear_model import LogisticRegression
import pickle

clf = LogisticRegression(random_state=0).fit(X_train, y_train)
filename = 'finalized_model.sav'
pickle.dump(clf, open(filename, 'wb'))
y_pred_lr=clf.predict(X_test)
print(y_pred_lr)
```

c.

```
[10] from sklearn.metrics import confusion_matrix,accuracy_score

     score=accuracy_score(y_test,y_pred_lr)
     print(score)
```

```
0.7673357664233577
```

```
[11] from sklearn.metrics import f1_score

     f1=f1_score(y_test, y_pred_lr, average='macro')
     print(f1)
```

```
0.7590704194026415
```

```
[12] from sklearn.metrics import confusion_matrix

     cm=confusion_matrix(y_test,y_pred_lr)
     print(cm)
```

```
[[319 153]
 [102 522]]
```

```
from sklearn.metrics import precision_score

prec=precision_score(y_test_test,y_pred_pred)
print(prec)
```

```
0.7733333333333333
```

```
from sklearn.metrics import recall_score

rs=recall_score(y_test_test,y_pred_pred)
print(rs)
```

```
0.8365384615384616
```

d.

fuck that downvote them for stalking you in the future like when that car in front of you makes all the turns you re about to make
i did the all the time with my wife texting is a waist of time if we are nt doing something together i have better shit to do with my time
i wonder how many accidents have been caused by this i ve come close a few times myself in the history of mankind i m willing to wager that millions of men have
who got laid off him or his colleague if the other guy still works there your friend knows where the guy works and can get ahold of him there plus since your frie
the first time i cleaned restrooms at my new job there was a massive pile of shit on the seat and floor of a woman s toilet it was also splattered all over the sink f

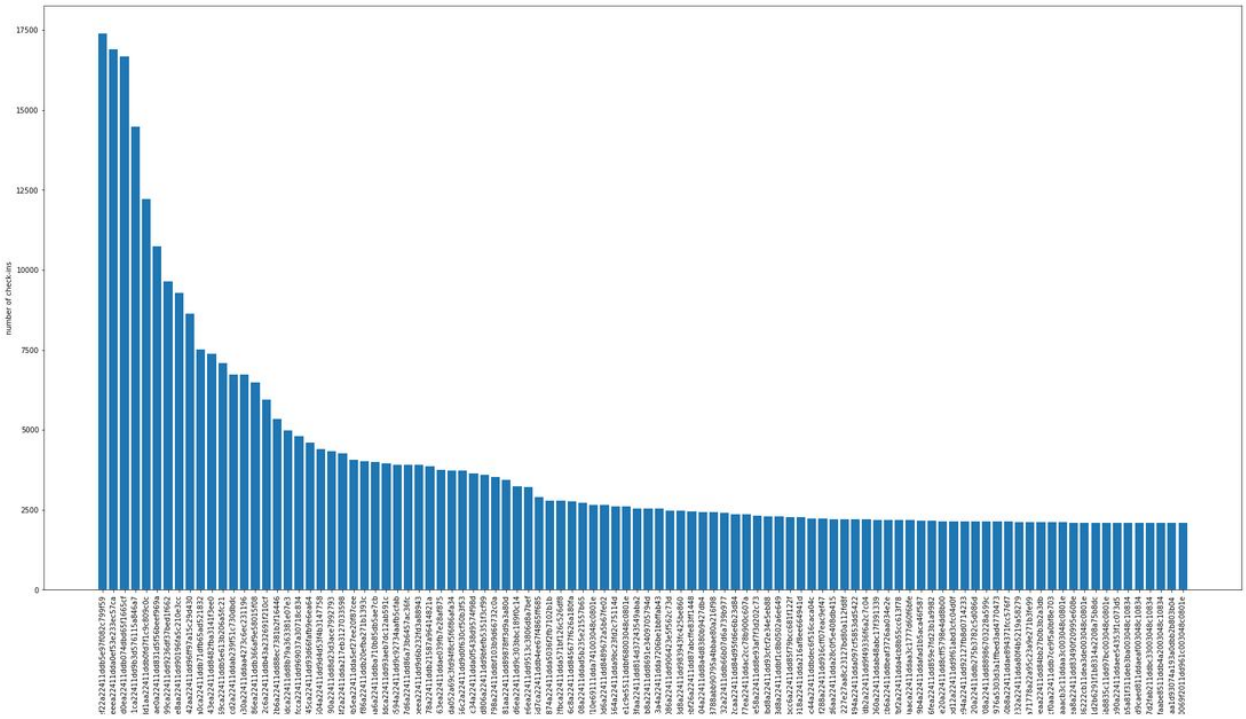
These are the comments that do not perform well

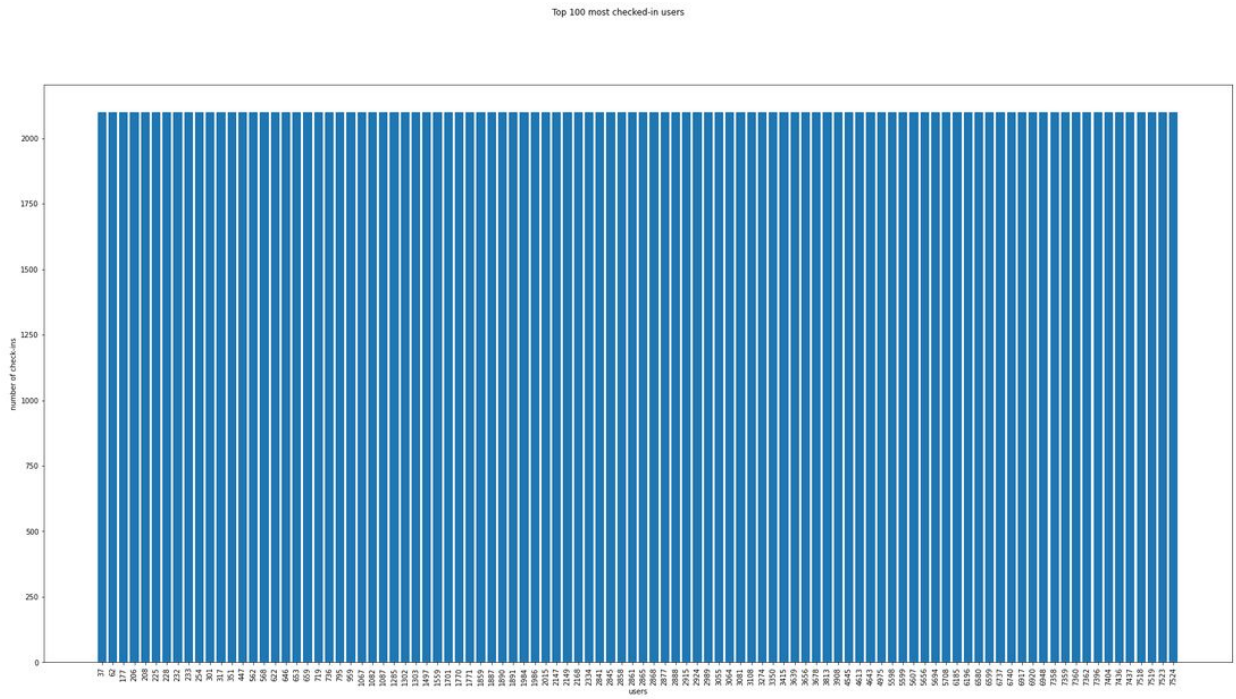
```
it took me a while to even get my legally required health insurance because  apparently  the government did nt know i exist they could nt find anything about me in t
it would be funny if she was also a redditor and gives op a big fat f using that same marker
who gives a flying fuck about columbine anymore  it was a fucking life time ago even the edgey cunts who thought the shooters were just cute and misunderstood have m
he probably sold it for drug money why ca nt he contact his former colleague  i call bullshit on that let s just say if it was my game he would be forced to contact
it s a bit dumbed down  but technically  by freeing up the ram  and killing off nonessential processes  you should have a more efficient power usage  not really a fa
```

These are comments which did perform well

The comments that do not perform well are due to misclassification which means that the problem arises when a humour comment uses words that are more commonly observed in news or vice versa. The assumption that I am making here is that duplicates had to be removed because otherwise I was getting much higher scores when there was duplication (approximately a 25 percent difference) which is obvious. The comments that performed well were news comments which used vocabulary and structures similar to most news comments and humour comments similar to most humour comments.

Section - II
Q1. a)





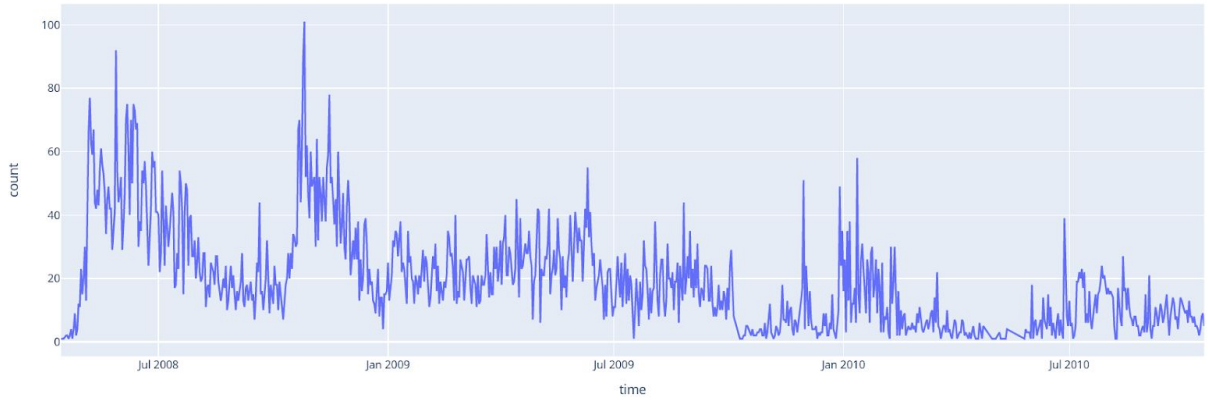
The most popular locations have a high density of check-ins, this is with respect to the check in of people. If we observe the graph, it is visible that the curve for most check ins has a steep drop which shows that the most popular locations have a lot of population in them and thus the location to people ratio is very skewed. We can observe that for the users checking in, the most checked in users are extremely uniform and this curve takes a very subtle decline. Thus combining these 2 insights, we can conclude that most people have a similar amount of check ins but they all visit the most popular locations disproportionately to other locations. Here I have only put the graphs for 100 users and locations because more entries made the labels not visible properly.

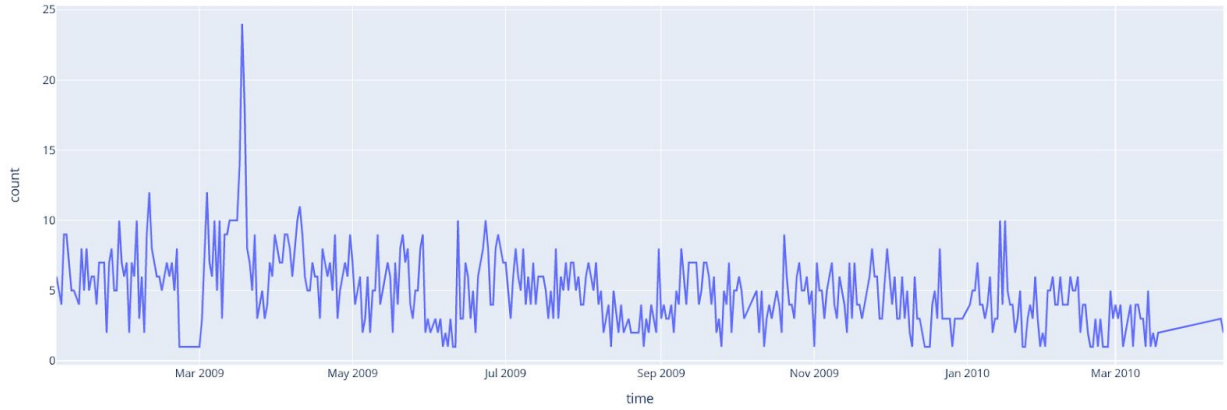
b)

We can observe that most of the heat is situated in and around the USA and some in certain parts of Europe. This is despite that fat that Asia has a much higher population than these other continents. This could be either due to the lack of access to ICTs in and around 2008-10 in other parts of the world (Ubiquitous computing was not really popular back then which have a much deeper global penetration now), it could also be down to better privacy policies and laws in place in other areas, it could be because of the popularity of locations based social

networking provided in the global north, and in general more of a culture that facilitated such check ins. Here I have plotted 50,000 data points due to computational constraints.

c)





Locations have a periodic spike around June in terms of check ins. Thus we can use this temporal information to advertise locations to people around this time of the year. This in combination with data about where the users are mostly situated gives us a good idea into what type of advertising can be done because we can make investigations about the cultures in the regions with the most heat and target them specifically. This also gives a good insight into what languages need to be used while advertising certain locations. The most popular locations can also be used as a competitive template and thus insights into attracting more people can also be derived.

Q2. The information in this dataset can be leveraged in the following ways:

- Getting the correct geospatial and time insights to better make targeted advertisements.
- These insights can help to track criminal activity and be used for law enforcement.
- This information can be used for disaster management and thus people in trouble can be found more easily and evacuation etc can also be tracked.
- This dataset can be used to evaluate disease spread and thus keep things under control. (Similar to Arogya Setu)

- This information can be used to better plan tourism policies and gain insights into checking in behaviour and understanding what attracts people in specific locations.

Some of the challenges in terms of privacy and security

- These insights can reveal sensitive information about a person's whereabouts thus leading to real life consequences like stalking, blackmailing and other forms of abuse.
- It is a violation of privacy because it can help the government to track down protests and other forms of dissent and imprison them.
- This information may create a distrust between employers and employees because of constant tracking. One such example is observed in India where Aasha workers don't trust medical professionals because they are scared of being tracked.
- This information can be sold to rogue third parties which can result in terrorism and hate crime.
- It can restrict a person's personal freedom and thus creating sense of paranoia and anxiety which can result in various mental health issues.

Section III
Q1. a.

```
print('The mean of Ratcliff Obershelp (sequence based)similarity on Twitter-Facebook is '+str(statistics.mean(simTwFaRat)))
print('The standard deviation of Ratcliff Obershelp (sequence based)similarity on Twitter-Facebook is '+str(statistics.stdev(simTwFaRat)))
print('The mean of Jaro Winkler (edit distance based) similarity on Twitter-Facebook is '+str(statistics.mean(simTwFaJaro)))
print('The standard deviation of Jaro Winkler (edit distance based) similarity on Twitter-Facebook is '+str(statistics.stdev(simTwFaJaro)))
```

```
The mean of Ratcliff Obershelp (sequence based)similarity on Twitter-Facebook is 0.700988165113022
The standard deviation of Ratcliff Obershelp (sequence based)similarity on Twitter-Facebook is 0.3391560783650944
The mean of Jaro Winkler (edit distance based) similarity on Twitter-Facebook is 0.7931042250027888
The standard deviation of Jaro Winkler (edit distance based) similarity on Twitter-Facebook is 0.2663221999961097
```

```
print('The mean of Ratcliff Obershelp (sequence based)similarity on Facebook-Instagram is '+str(statistics.mean(simFaInRat)))
print('The standard deviation of Ratcliff Obershelp (sequence based)similarity on Facebook-Instagram is '+str(statistics.stdev(simFaInRat)))
print('The mean of Jaro Winkler (edit distance based) similarity on Facebook-Instagram is '+str(statistics.mean(simFaInJaro)))
print('The standard deviation of Jaro Winkler (edit distance based) similarity on Facebook-Instagram is '+str(statistics.stdev(simFaInJaro)))
```

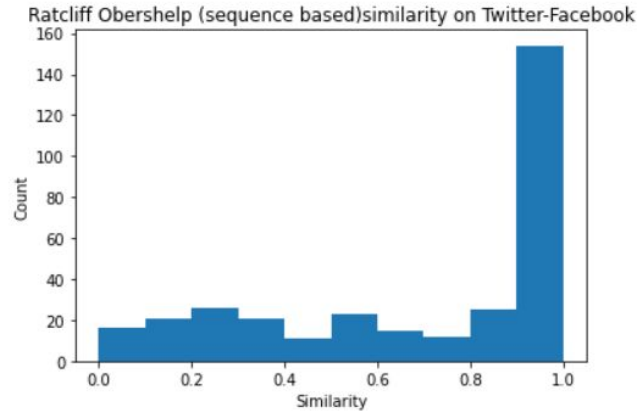
```
The mean of Ratcliff Obershelp (sequence based)similarity on Facebook-Instagram is 0.7206362511972856
The standard deviation of Ratcliff Obershelp (sequence based)similarity on Facebook-Instagram is 0.3279176755494043
The mean of Jaro Winkler (edit distance based) similarity on Facebook-Instagram is 0.8067161922854186
The standard deviation of Jaro Winkler (edit distance based) similarity on Facebook-Instagram is 0.2611728938309347
```

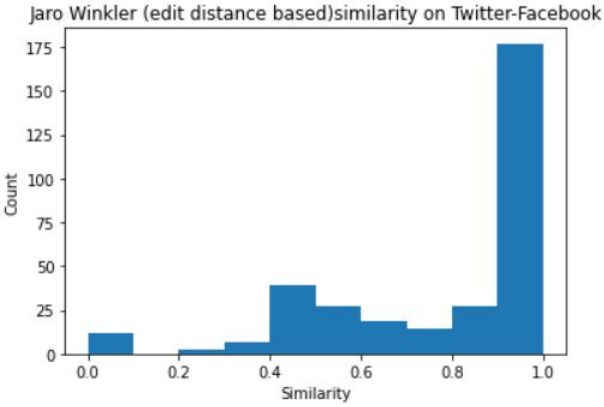
```
print('The mean of Ratcliff Obershelp (sequence based)similarity on Twitter-Instagram is '+str(statistics.mean(simTwInRat)))
print('The standard deviation of Ratcliff Obershelp (sequence based)similarity on Twitter-Instagram is '+str(statistics.stdev(simTwInRat)))
print('The mean of Jaro Winkler (edit distance based) similarity on Twitter-Instagram is '+str(statistics.mean(simTwInJaro)))
print('The standard deviation of Jaro Winkler (edit distance based) similarity on Twitter-Instagram is '+str(statistics.stdev(simTwInJaro)))
```

```
The mean of Ratcliff Obershelp (sequence based)similarity on Twitter-Instagram is 0.8363071901277543
The standard deviation of Ratcliff Obershelp (sequence based)similarity on Twitter-Instagram is 0.2945499470775639
The mean of Jaro Winkler (edit distance based) similarity on Twitter-Instagram is 0.8833473676361294
The standard deviation of Jaro Winkler (edit distance based) similarity on Twitter-Instagram is 0.2316370553971256
```

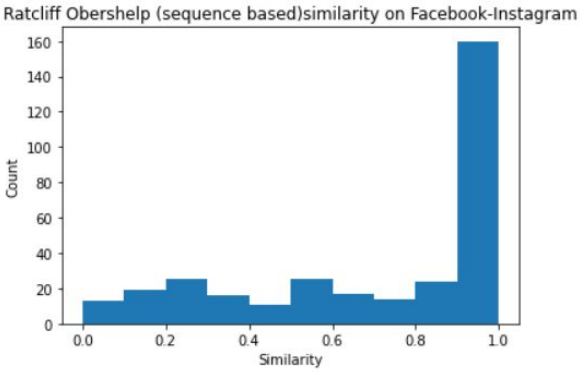
I have used 2 different types of distance metrics over here, namely Ratcliff Obershelp which is a longest common subsequence matching and I have used the Jaro Winkler similarity which is based on edit distance. Both of these show similarity across all the comparisons. However, in all 3 platforms we can see that the Jaro Winkler similarity based on edit distance performs better than the Ratcliff Obershelp in most cases given that the mean is higher, along with a lower standard deviation. I will not make distinctions between each of the 3 cases because the same trends are observed but if I had to choose the best metric, I would perhaps use a combination of all 3, edit distance and common subsequence similarity along with the third, ie. token based similarity to come up with the final similarity. All 3 have their own merits and demerits and thus this could be helpful. However, if told to select one, it will be the Jaro Winkler similarity.

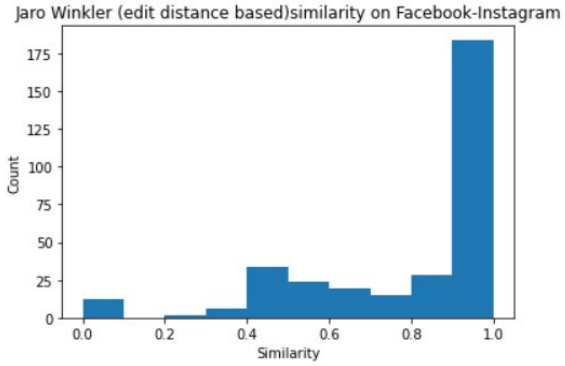
b. i)


Ratcliff Obershelp (sequence based)similarity on Twitter-Facebook
Count
Similarity
0
20
40
60
80
100
120
140
160
0.0
0.2
0.4
0.6
0.8
1.0

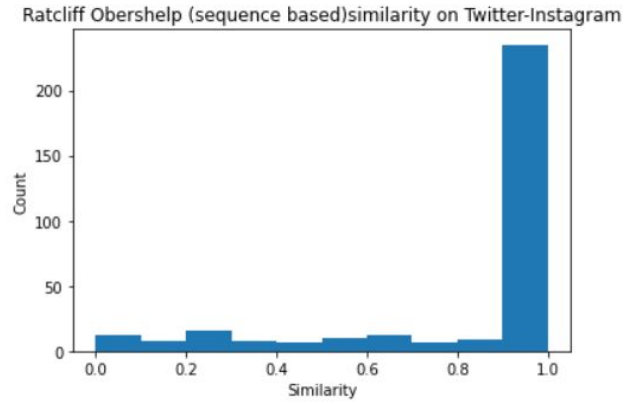

Jaro Winkler (edit distance based)similarity on Twitter-Facebook
Count
Similarity
0
25
50
75
100
125
150
175
0.0
0.2
0.4
0.6
0.8
1.0

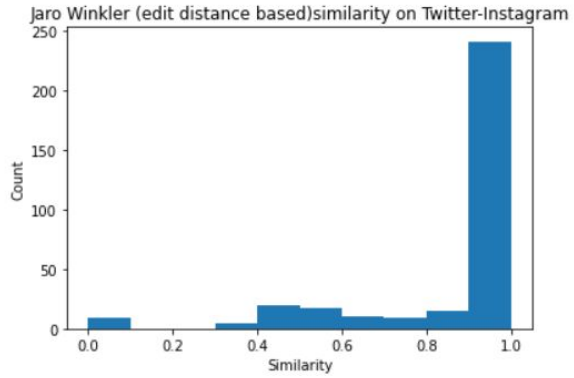
ii)


Ratcliff Obershelp (sequence based)similarity on Facebook-Instagram
Count
Similarity
0
20
40
60
80
100
120
140
160
0.0
0.2
0.4
0.6
0.8
1.0



iii)





From these plots I can infer that the resolution between the pairs of social media platforms is very much possible. However, I have not tested for two distinct people identities with distinct handels. Thus, testing for that can provide what is the best distance metric. From the plots it is clear that Jaro Winkler performs better in all 3 platforms. We can thus create identities for the same person using different social media platforms and leverage this in various ways like gaining data from all three platforms, hacking into all 3 platforms etc.

Section - IV

Q1. If I were to take up a project that wants to predict aadhar number in India, the steps that I will take is:

- Scope for available datasets of aadhar details

- Try to establish a correlation between aadhar number and demographic details of people. This can be done through running a regression for every single digit in Aadhar numbers with various demographic details like location, date of birth, phone number etc.
- Once the correlation is established, we can reasonably predict trends by using a machine learning algorithm that can provide accurate numbers depending on the demographics of a person.
- So, with a given confidence and error we will be able to predict/ estimate aadhar numbers for any given person.

But the problem that arises here is that from my research about aadhar numbers, they seem to be random or pseudo random (first 11 digits are random, 12th digit is a checksum of the previous 11 digits), thus there would not be a reasonable correlation of statistical significance. I have said here that they are pseudo random because there are other factors like resalting, uniqueness checks and rejecting special numbers, based on nothing but first come first serve. Thus, with this knowledge I think it would not be possible for me to accurately predict aadhar numbers of people in India. Another innovative method that I can think of is using other government identification which is not as robustly designed or has numbers depending on demographic details can be used instead like the driving license. This can potentially be back tracked to Aadhaar numbers.

Q2. The use limitation principle is difficult to implement because:

- The user experience provided to each user is made extremely personalisable and subjective to that user, to achieve this more and more data from the user is required. Thus creating better experience for their target stakeholders is essential. This makes it difficult to implement the use limitation principle.
- A large part of industry comes from selling data to third party advertisements and thus extensive data about the user is necessary to create targeted ads for particular users. Targeting ads leads to better sales thus affecting profits directly. Thus there is a direct profit in not implementing use limitation to full extent.
- There are a lot of third party application integrations in various platforms which can make it difficult for the platform to control the data acquired by these applications. One such example is the cambridge analytica scandal.
- Another problem that occurs is giving control to the user. The user may not fully understand what setting amounts to what directly and thus this requires better ways of self disclosure. As we saw from trends in facebook, giving the user control backfired monetarily and thus new defaults were established. The users tend to stick to defaults and thus this was a major limitation.
- There is a general lack of awareness about the privacy and security in social media in the general public and this knowledge gap creates a whole host of assumptions. To better reflect the use limitation principle, better ways of portraying privacy policies need to be figured out than a long document which basically no one can go through.

Q3. If I were working in one of these organisations, to defend the point, I'll use the following arguments:

- The acquired data can further help in strengthening the development and design of better security solutions for the users.
- The collected data can create more personalizable experiences for each user with better product, groups, content, friends and activities recommendations.
- The data collected is anonymised and is stripped off of any personally identifiable information, with sensitive data being treated carefully thus making it more secure and increasing privacy of data.
- Situating the company to better understand consumer's attitude and expectations about data.
- Incorporating user data features to help law enforcement and create better safety features like tracking criminals, emergency requests among others. Malicious activity can also be more easily tracked.