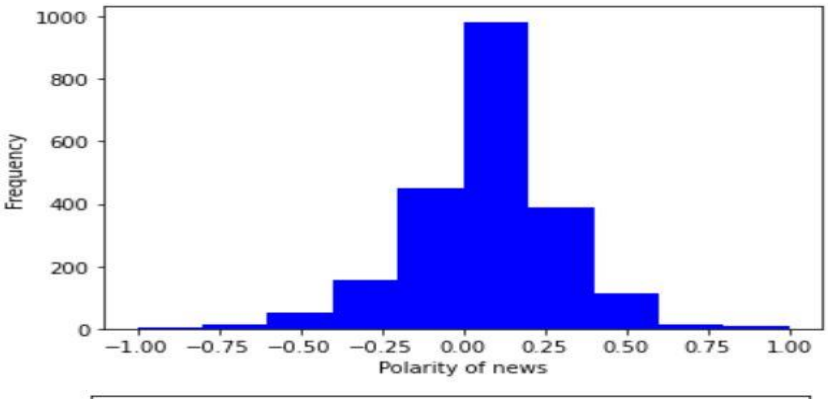
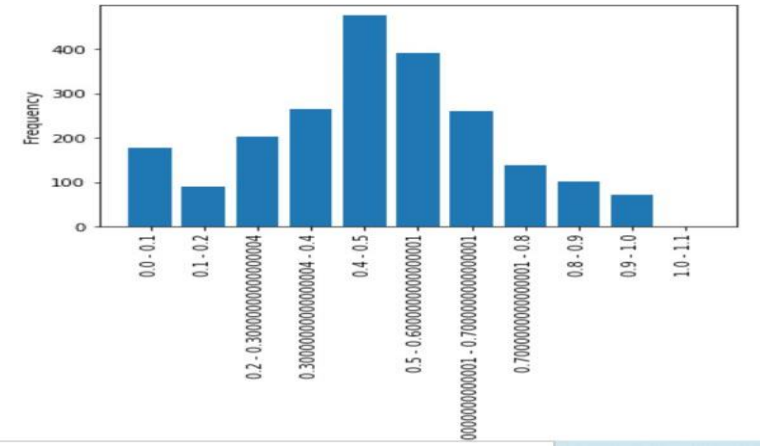
Section 1

1)a)

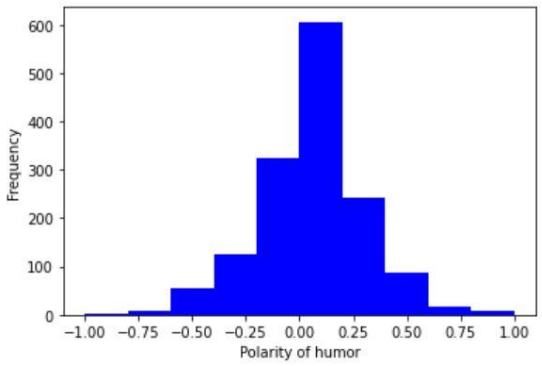
Polarity of news:



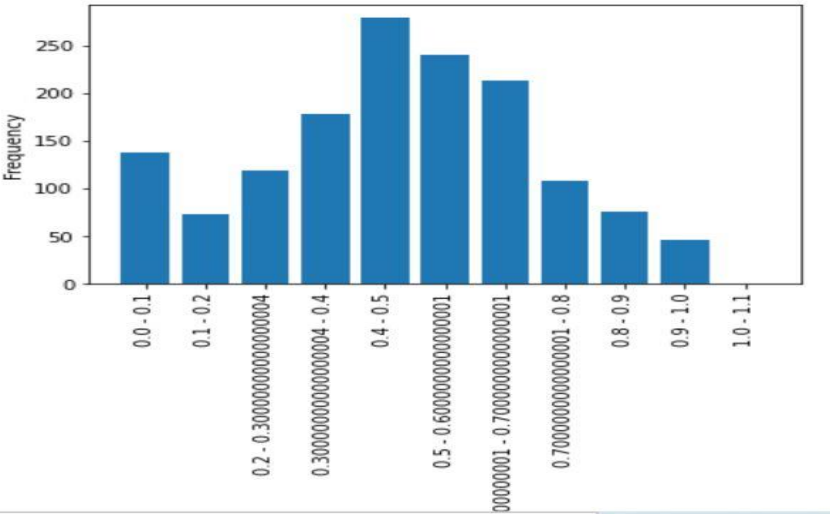
Subjectivity of news



Polarity of humor:



Subjectivity of humor:



Inference:

Subjectivity and polarity of most of the words lies in the median hence we can say most of the words are neutral(neither too subjective-less subjective or positive polarity-negative polarity)

Standard deviation and mean of the above:

```
Mean of polarity humor is:  0.05158728575619279
Standard Deviation of polarity humor is:  0.24866391452327927
Mean of polarity news is:  0.064048534801197
Standard Deviation of polarity news is:  0.23041556585535447
Mean of subjectivity humor is:  0.47140332998753287
Standard Deviation of subjectivity humor is:  0.24269324821277793
Mean of subjectivity news is:  0.4704749753548112
Standard Deviation of subjectivity news is:  0.23274569397299044
```

Inference:
Mean of polarity of humor is very less than that of polarity of news, subjectivity of both news and humor hence we can say when someone talks in humor then polarity is very less as compared to other cases.

b)

The following are the cases where my model predicted well:

- my husband would have done the exact same thing you keep pulling shit like that your wife wo nt try to do anything sexy anymore--→ Humor
  This cannot be a news because of words like “shit and sexy” and the start of this comment is ‘my husband’ which sounds like an opinion.
- if the late justice had died on january 15th of next year ok they might have a point at that point it s a week until inauguration nothing s going to get done in that time but it would be classy to consult with his successor to get the process started a week sooner but cmon guysinauguration is 11 12 months away insisting on it now is just plain childish-→News
  This is classified as news as it have words with more subjectivity and doesn’t have first person word which gives an idea of news.
- literally the reason i ve only done digital since the new systems came out friend s husband borrowed skyrim then magically ca nt find it 3 weeks later when he starts moving because of divorce-→ Humor
  This is classified as humor because of using first person hence it sound like an opinion instead of news.
- the only people using the term firewall are in the media usually in a way that is negative towards hrc s campaign and yet redditors jump on the this is offensive bandwagon lol wtf -→News
  This is classified as news as well because of words with more subjectivity.
- it s important for the government to provide a consistent and cohesive narrative if they are to help squash future acts of rebellion cliven bundy and his ilk are lawless violent redneck anarchists and none of the things they protested had any validity whatsoever→News
  This is classified as news because there have been worlds with more subjectivity and no singular words.

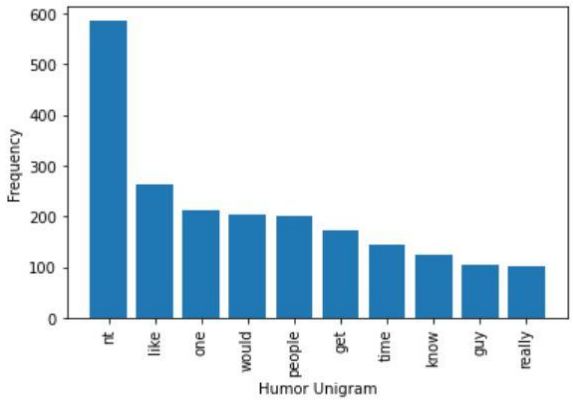
The following are the cases where my model failed:

- most importantly inventory does nt track preferences almost as important stocktaking only tells you about the past whereas prices tell you about the present and people s beliefs about the future→Predicted as Humor but actually News
  This is predicted as humor because of use of second person such as “you” which made it predicted as humor.
- a railroad engineer must be sure not to lose his train of thought or he might go down the wrong track →Predicted news but is engineer.
  Words like railroad engineer, he, wrong track can be part of a news and there’s no first person which indicated it to be humor hence predicted as news.
- i do nt understand i thought women could nt do this maybe those college classes on how not to rape should include women → Predicted as humor but is news
  This is classified as humor because of use of first person which indicates the model it to be an opinion.
- he s scum but he s smart he does nt just send 15 million to someone claiming to be kanye s boy it s not like buying something of ebay he s got lawyers and financial

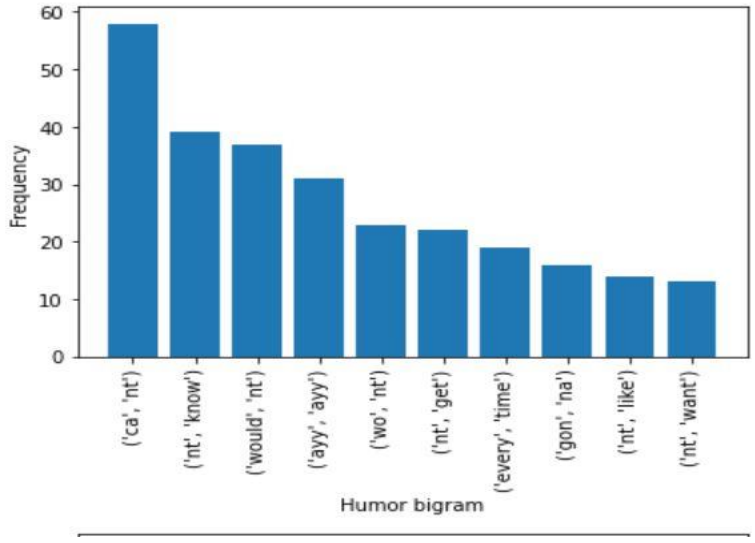
advisers and all sorts of shit he s trolling people either for the lulz or maybe as ugreypoweroz suggests to hide money this is fucking ridiculous and people are just eating it up cause it sounds like a funny thing to happen → Predicted as news but is humor
This is happened because of use of second person and more subjective words.

- cool all you need to do is inject stem cells onto the printed copy implant it back into the patient and presto no need to take antirejection meds for the rest of your life→Predicted as humor but is news.
Use of first person leads this to be predicted as humor but it is news.

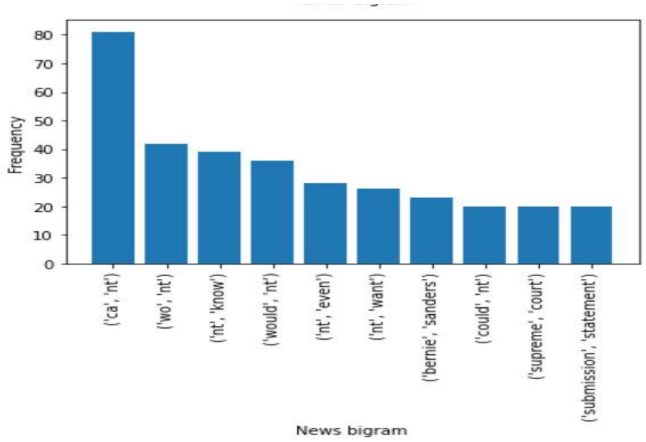
Humor Unigram:



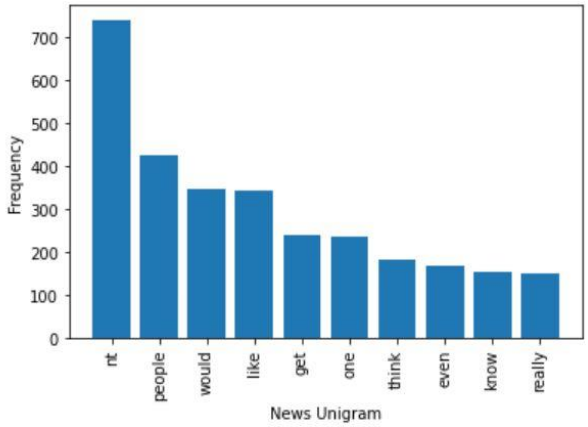
Humor bigram:



News bigram:



News Unigram:



Inferences: Number of nt for news unigram are really high as compared to other words in unigram and humor unigram.
News contains words with more subjectivity such as: Supreme court, submission, statement as compared to humor which have words with good polarity but not good subjectivity.

2)

a) Following are the preprocessing and splitting data and modelling.
I have used randomizing the data for splitting for better ML classifier. I have used Logistic regression since I did hit and trial with SVM, random forest, Naïve bayes and got better accuracy in logistic regression.

-> Splitting features and label in two different nparray
-> Changing 1 where it's written 'humor' and 0 where it's labelled as 'news'.
-> Splitting dataset to training and test set by train_test_split in sklearn library.
-> Do vectorization with the help of CountVectorizer();
-> Make logistic regression model and apply it to X_train and y_train.
->After that test it with y_test to get prediction array called y_pred

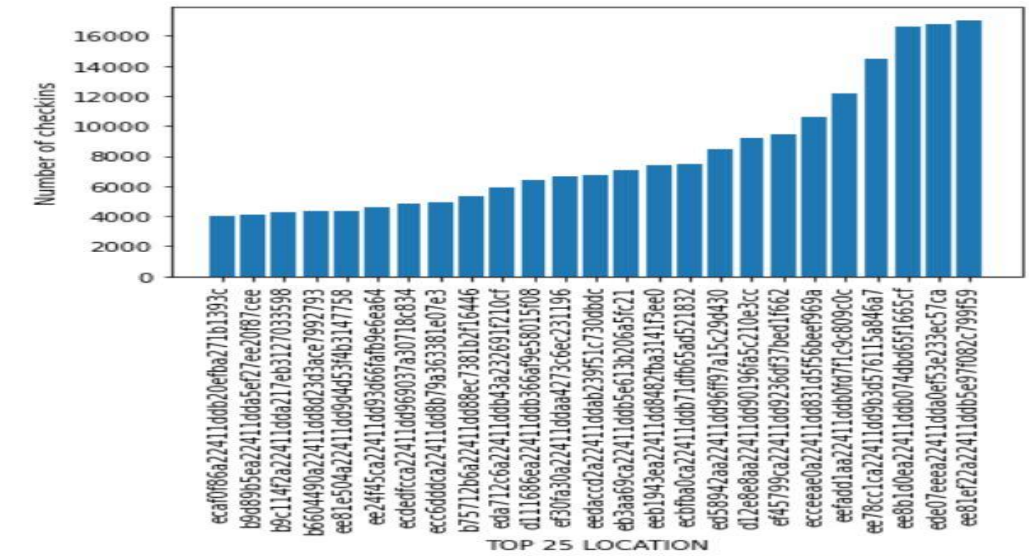
```
Xdf=pd.DataFrame(data=X, columns=['Polarity', 'Subjectivity']);
Ydf=pd.DataFrame(data=y, columns=['label']);
## splitting test and train dataset
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(features, label, test_size = 0.3, random_state = 1)
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score, classification_report
#making vection
vectorizer = CountVectorizer()
X_train = vectorizer.fit_transform(X_train)
X_test = vectorizer.transform(X_test)
#building classifier
classifier = LogisticRegression(random_state =0)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
print(accuracy_score(y_test, y_pred))

print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test,y_pred))
```

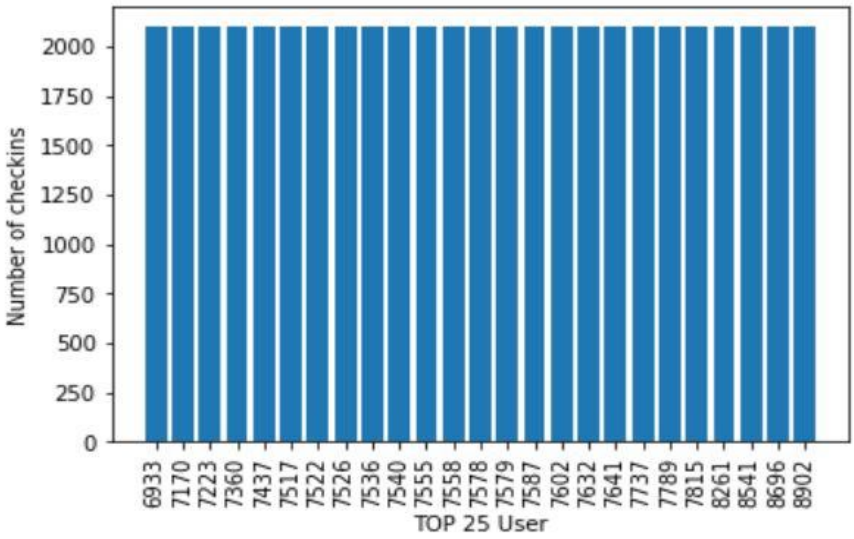
Following shows the given metrics respectively such as:
1)Accuracy score
2) Confusion Metrix
3) Classification report which contains precision, f1-score, recall

```
0.7928832116788321
[[545  84]
 [143 324]]
              precision    recall  f1-score   support

           0       0.79      0.87      0.83       629
           1       0.79      0.69      0.74       467

    accuracy                           0.79      1096
   macro avg       0.79      0.78      0.78      1096
weighted avg       0.79      0.79      0.79      1096
```

Section 2

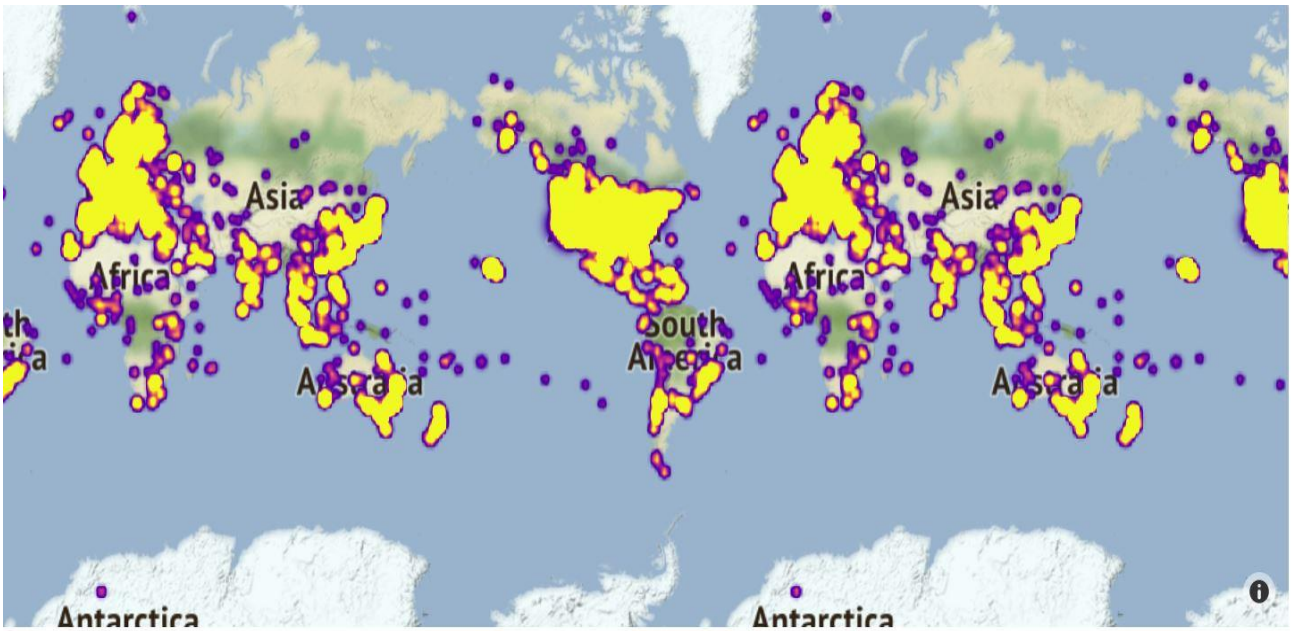
1)
a)
Plot Distribution of locations and users with number of check-ins:
Assumption: I have taken most frequent 25 users instead of plotting all numbers





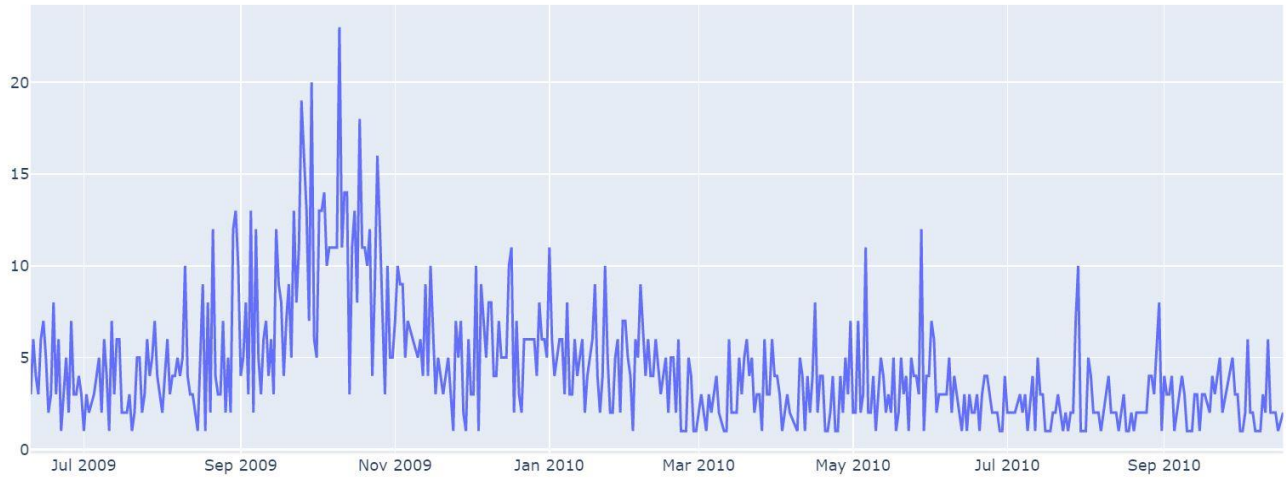
Inference:
Top 25 users have same number of checkins.
Out of top 25 location, the top three have identical number of checkins this can be interpreted as they could be close to eachother or they're really famous. Then the number of checkins are decreasing pretty much as they reached from 16000 to 4000.

b)
Displaying HeatMap

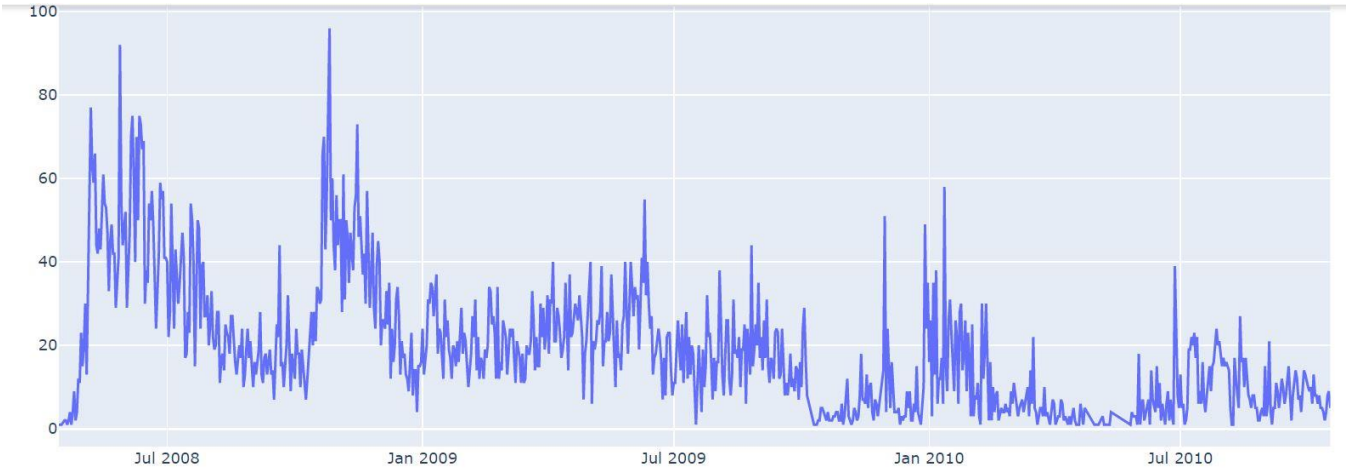


INFERENCE:
It is observed that North America and Europe have covered lot of heat in the heatmap which shows that many people visit Europe and North America as they both are tourist spots and there's really less checkins in polar regions such as Antarctica.

c) Time Series Graph of 1) Top User 2) Top location



The above is the time-series graph of top user. If we see a spike in user activity at a particular time then we can target that user by recommending him different different locations at that particular time so that we will visit that particular place.
Or this kind of information can be used to target an individual for meet at a given time when that particular person go to some locations if we see some pattern.



The above is the time-series graph of top location. We can infer from the given data that number of checkins are decreasing year by year in that particular location hence we can give discount or offers to the hotels in that location to increase checkins by advertisement.

b) Leverages:

If we have this dataset, we can do multiple things such as:

1) Tracking/tracing a user activity and then recommending him/her the types of hotels from our firm nearby his/her area.
2) We can get top k users and then show him/her advertisements regarding trips/hotels.
3) We can find out the hotels which have low check ins and can tell them offers like advertising for their own hotels.
4) We can recommend locations to third party who want to open hotels to tell what are the possible places where opening a hotel would be profitable with more checkins.

Challenges:

1) With the above dataset, we can see predict the user's location if we see a periodicity in his location.
2) One can harm user by detecting their location which is a breach to security.
3) Showing users advertisement corresponding to their likes is a breach in privacy of user since they recorded his/her likes.
4) We can breach information of people visited a particular hotel which might be against hotel's privacy policy.

Section 3

Metrics used: Damerau Levenshtein Distance and Edit Distance

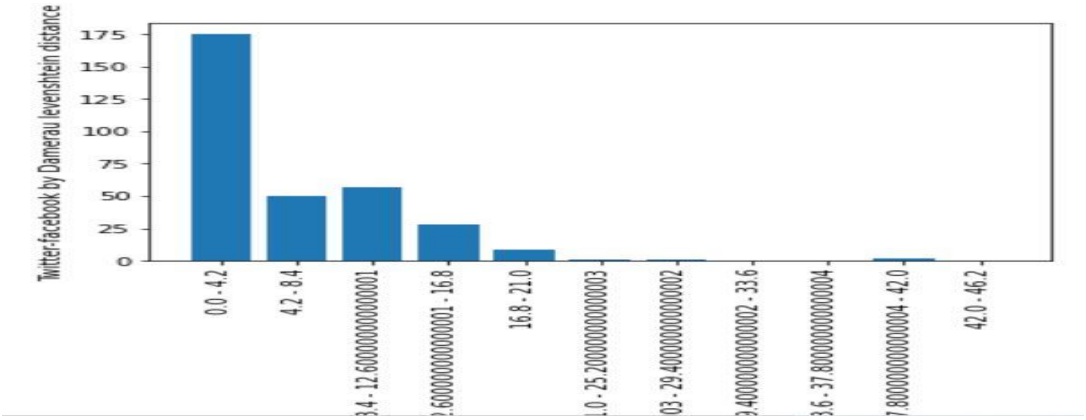
Inference:

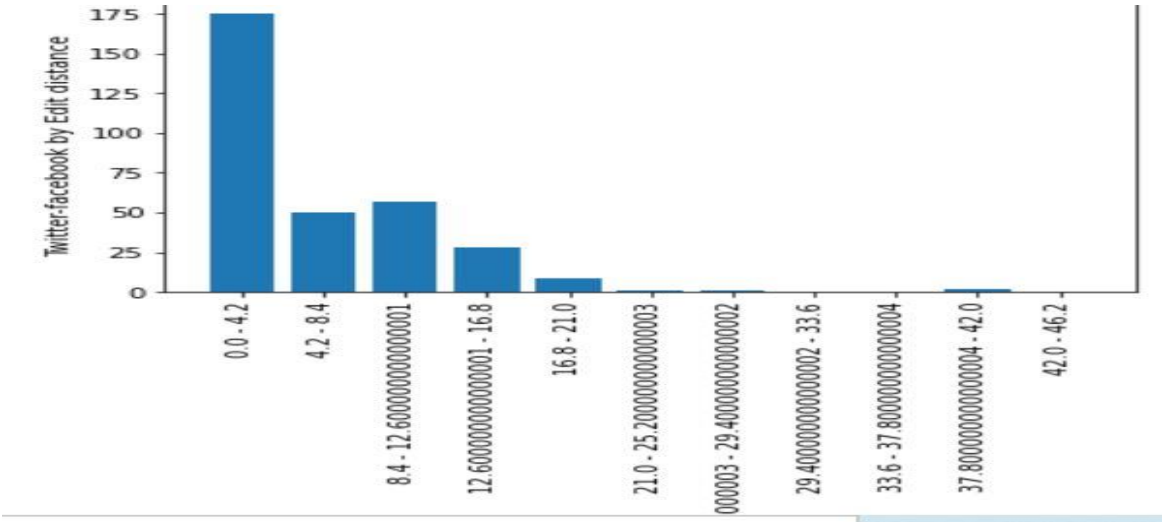
The Damerau–Levenshtein distance differs from the classical Levenshtein Distance by including transpositions among its allowable operations in addition to the three classical single-character edit operations (insertions, deletions and substitutions).
Edit Distance gives an minimum changes to tell how to convert from one string to another.
We can observe that Damerau-Levenshtein is also a subtype of edit distance but edit distance is better except for 4 cases.

Result:

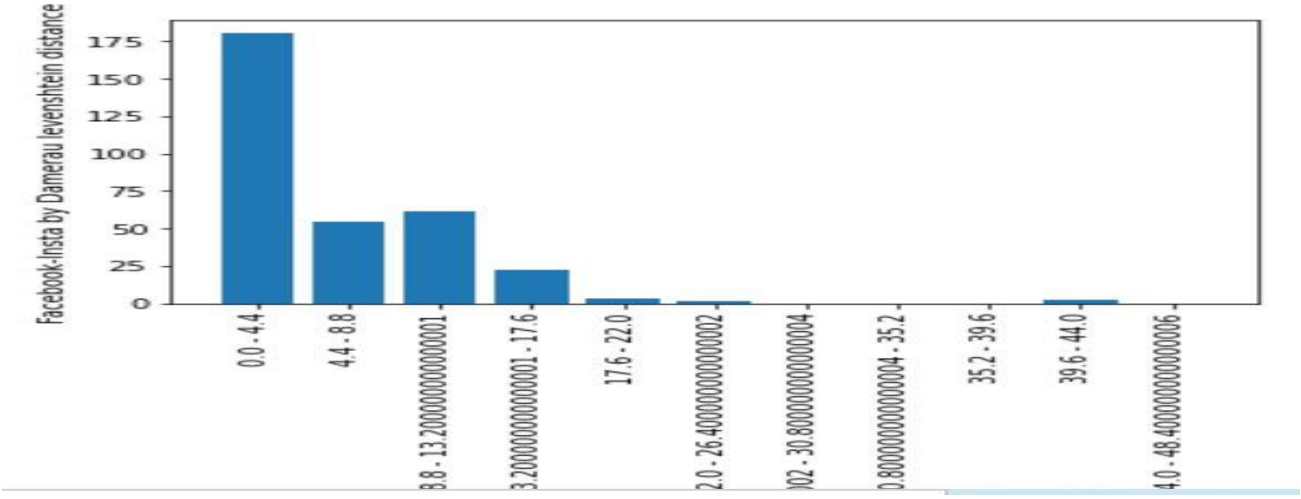
```
For twitter-facebook, better distance metric would be:  Edit Distance
For facebook-instagram, better distance metric would be:  Edit Distance
For instagram-twitter, better distance metric would be:  Edit Distance
```

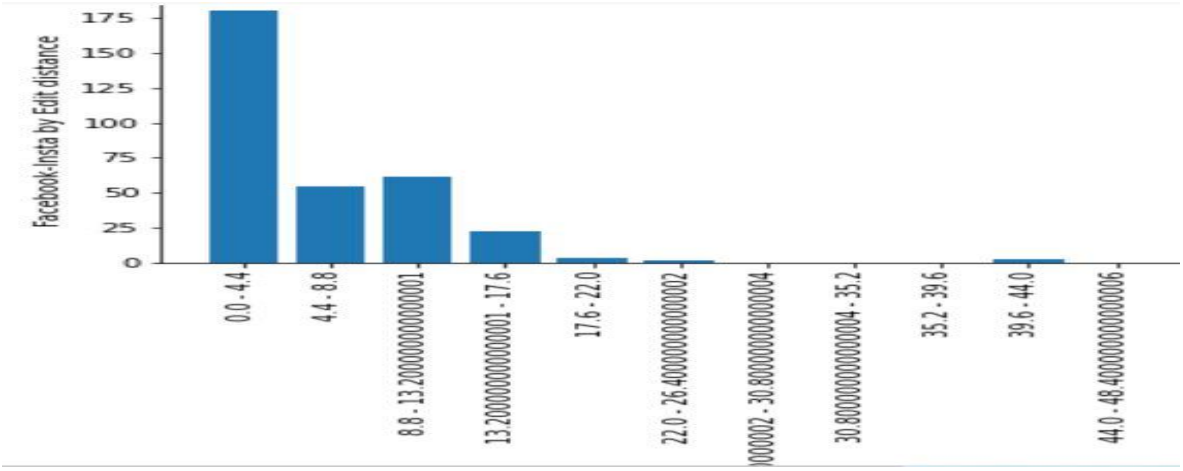
Twitter-Facebook metric by both Edit Distance and Damerau Levenshtein





Facebook-Instagram metric by both Edit Distance and Damerau Levenshtein





Section 4

Q1.

- SNN number can be found out using date of birth of users from different social media platforms. Then it will be used to predict SSN numbers.

- One can try out the same method to predict Aadhar number in India. For example, one can try out different information of a user such as date of birth, face etc information from social media and try to predict similarity if there any, from date of birth and face.
- It is not possible to predict Aadhar number in India as Aadhar number is a randomized number which doesn't have any relation from Date of birth or any other information in India. Hence it is not possible to predict Aadhar number.

Q2.
Use limitation principle: Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with purpose specification except: (a) with the consent of the data subject; or (b) by the authority of law.

It is hard to implement the 'Use limitation' principle due to following reasons:
Q2.
Use limitation principle: Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with purpose specification except: (a) with the consent of the data subject; or (b) by the authority of law.

- **Decreased Usability**: If we give users whole authority to choose between all the information then he/she will get lot of permissions/consent hence it will decrease the usability for that particular application.
- **Less control over 3rd party:** A company doesn't have total control over 3rd parties to which it will be giving the user's information as the third party might not follow it's guidelines.
- **Slow down of process of better application:** Since the company will wait for the user's consent, if the user doesn't give consent for longer period of time then the company have to wait before launching any update to that application.
- **Sensitivity of data is not clearly defined:** Every company defines the sensitivity of information on differently hence there's not much clear definition of what information a company will be considering for PII(Personal Identification Information).
  Revenue Loss: If the company doesn't share their data to third parties then they'll face a big loss in situation where user doesn't give consent to use it's information. For example: Whatsapp didn't get the expected profit hence it's privacy policies were a talking debate few months back.

Q3

Following are the reasons one can give to defend saving and using our information:

- Companies have recommendation algorithm i.e., they recommend and show us the information we are looking for or similar items like that.
  For example: If I like music on Youtube then Youtube will recommend me more music videos except some unnecessary videos which makes ease for me to explore more music.

- One can select the option of syncing their contacts with the social media platform and it tells if that person is on that network or not hence making our contacts from phone to our friends on social media as well hence contributing to increasing ways to connect an individual.
- Shows us relevant advertisements when we are looking for an example an item x, I get advertisement for item x and help us to shop in a better way.
- Saving information of users helps them to announce rewards for long term users or users who use their platform for longer period of time.
- Sharing information to third parties for better service such as amazon giving delivery details to the delivery company we have ordered our product.