CSE 648
Privacy and security on online social media
Mid-Sem report

Ajit Singh 2018009

Section 1

Q1

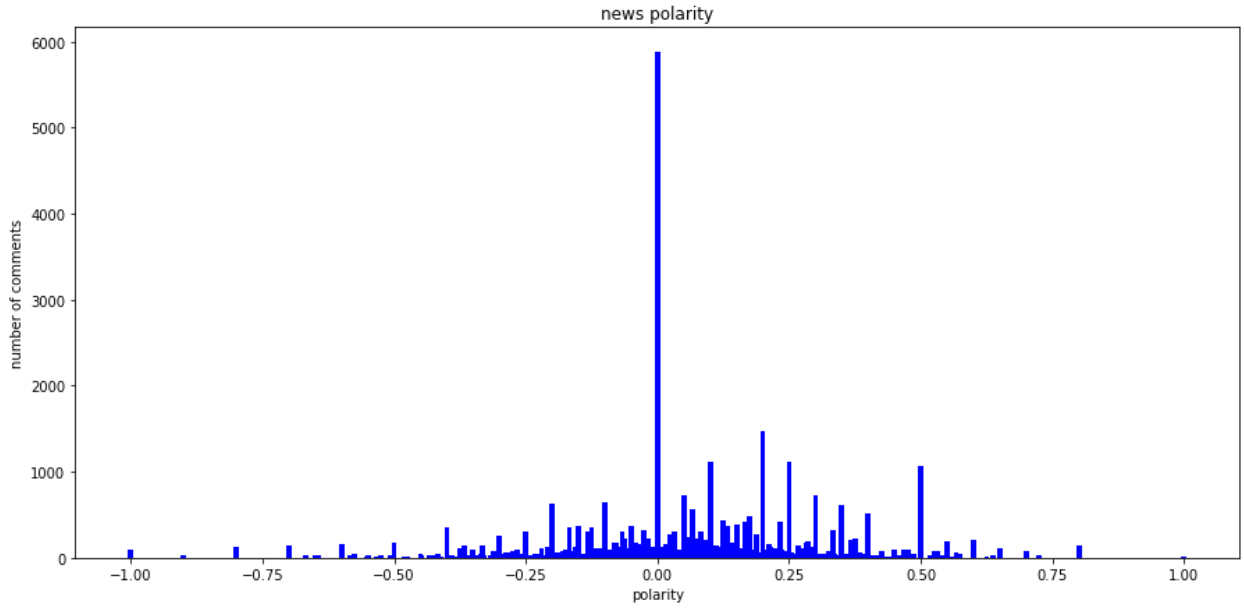
A

Preprocessing:

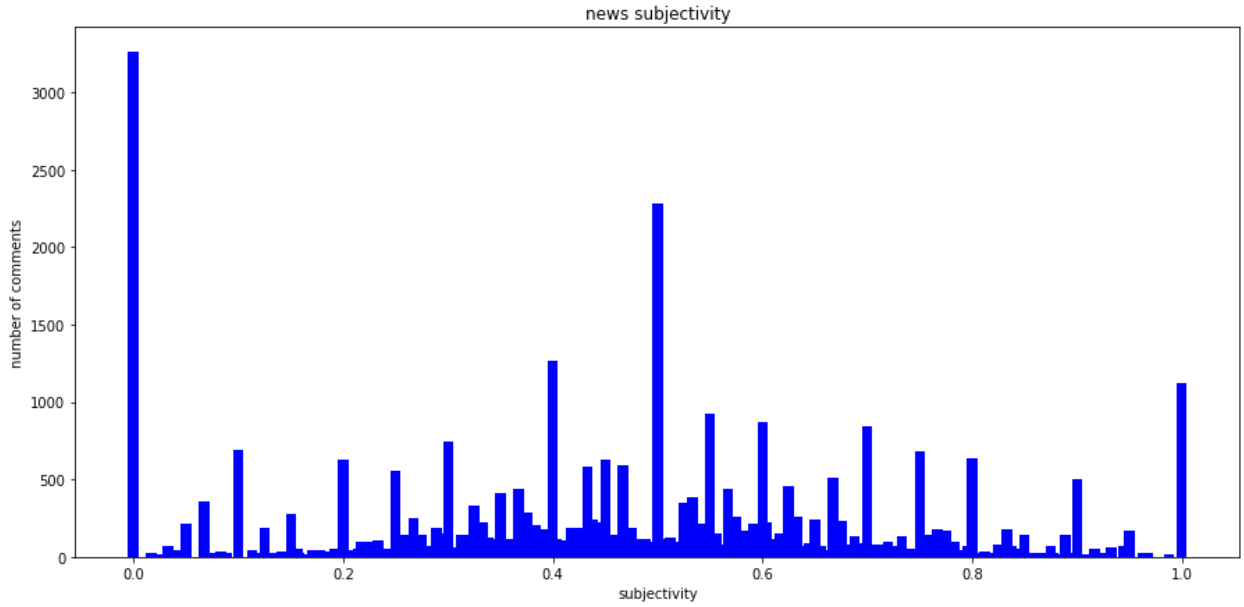
1. Removed duplicate entries from the dataset.
2. Removed words with lengths smaller than 3.
3. Removed URLs.
4. Removed stop words.

sentiment analysis library used: Textblob

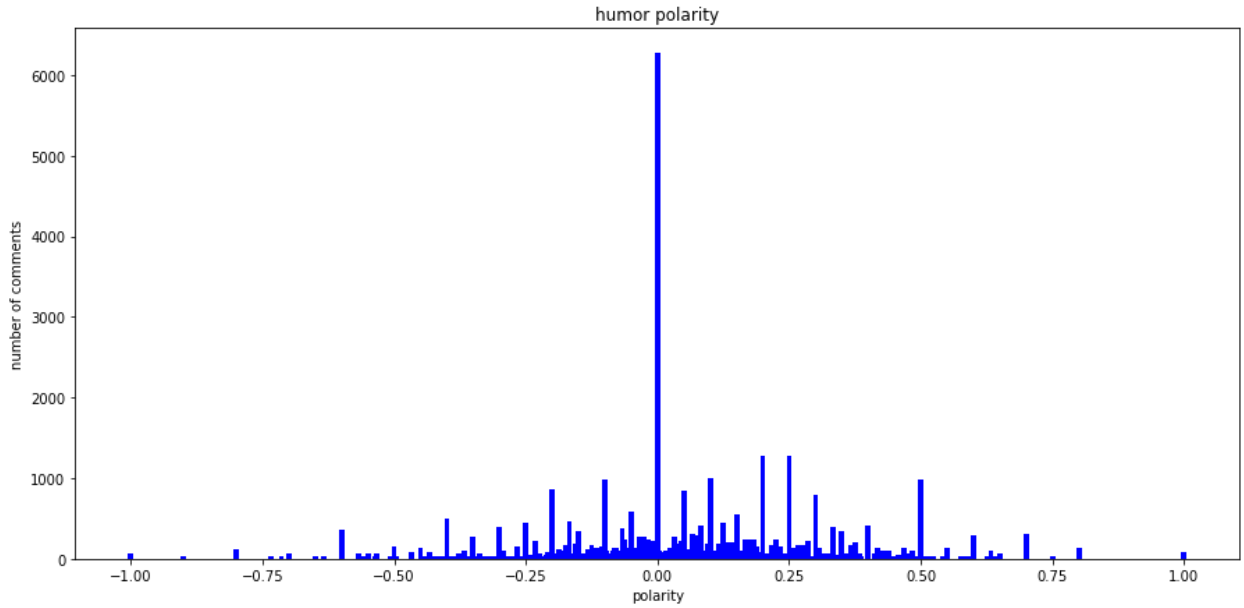
News polarity:
Mean 0.0639129303826589
Standard deviation 0.2303242080076376



News subjectivity:
Mean 0.4691590125773786
Standard deviation 0.23192148906041926

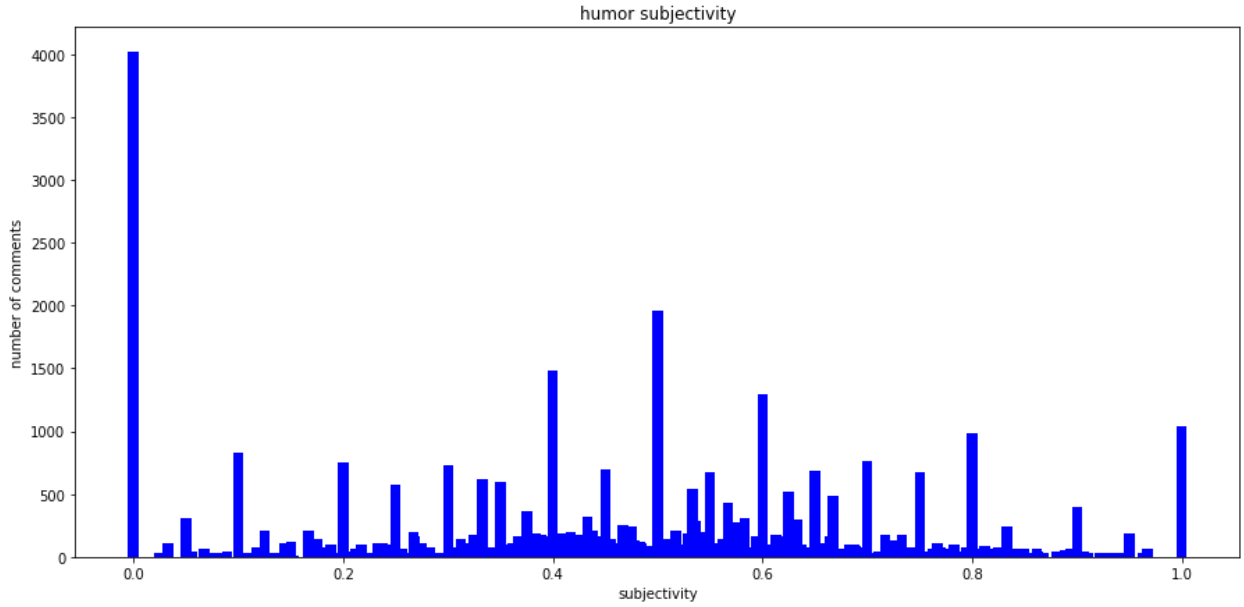


Humor polarity:



Mean:0.05231557130778092
Standard deviation: 0.24779379253602302

Humor subjectivity:



Mean: 0.47071756809233806
Standard deviation: 0.24243668370161633

Observations:

1. The polarity of the news and humor category are forming a sort of bell-shaped graph that denotes the polarity is on the neutral side
2. The polarity graphs are with a small tilt towards the positive side.
3. The subjectivity of the news and humor category is on the neutral side(objective) with a small tilt towards the positive side. (subjective).
4. It can be noted that the standard deviation is greater in the case of humor.

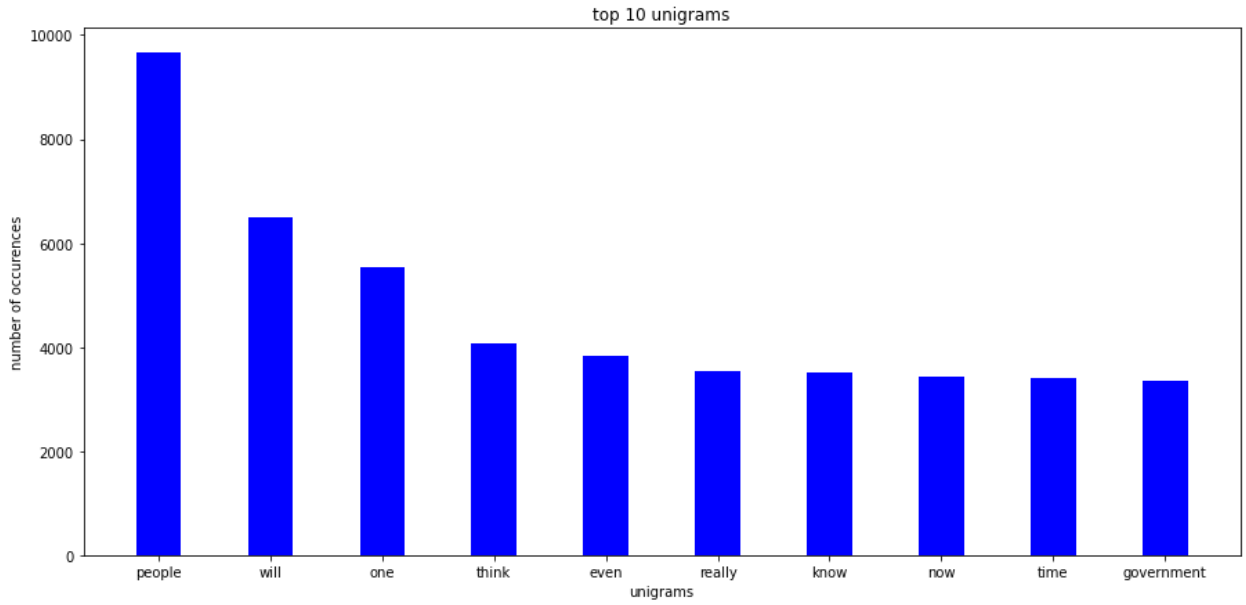
B

Stop words are removed from the text corpus so that ngrams only contain words that have some meaning.

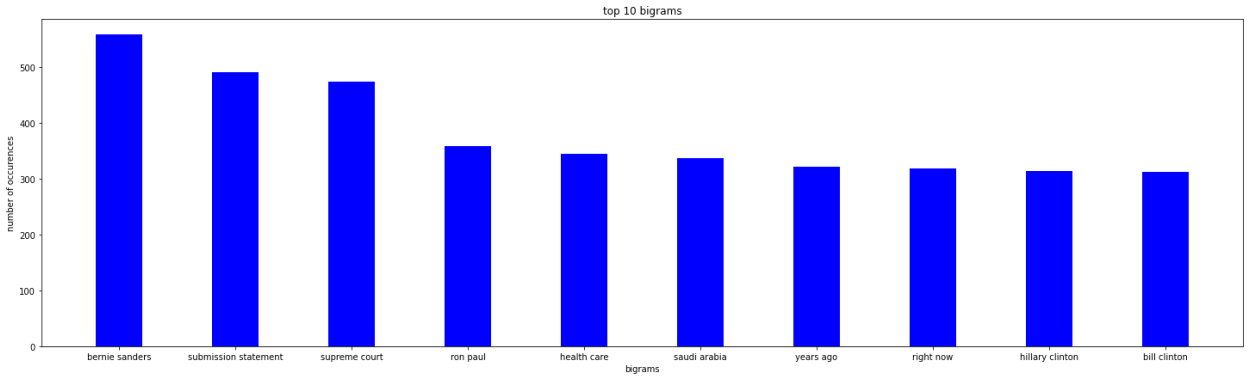
Library used:

Nltk.ngrams.

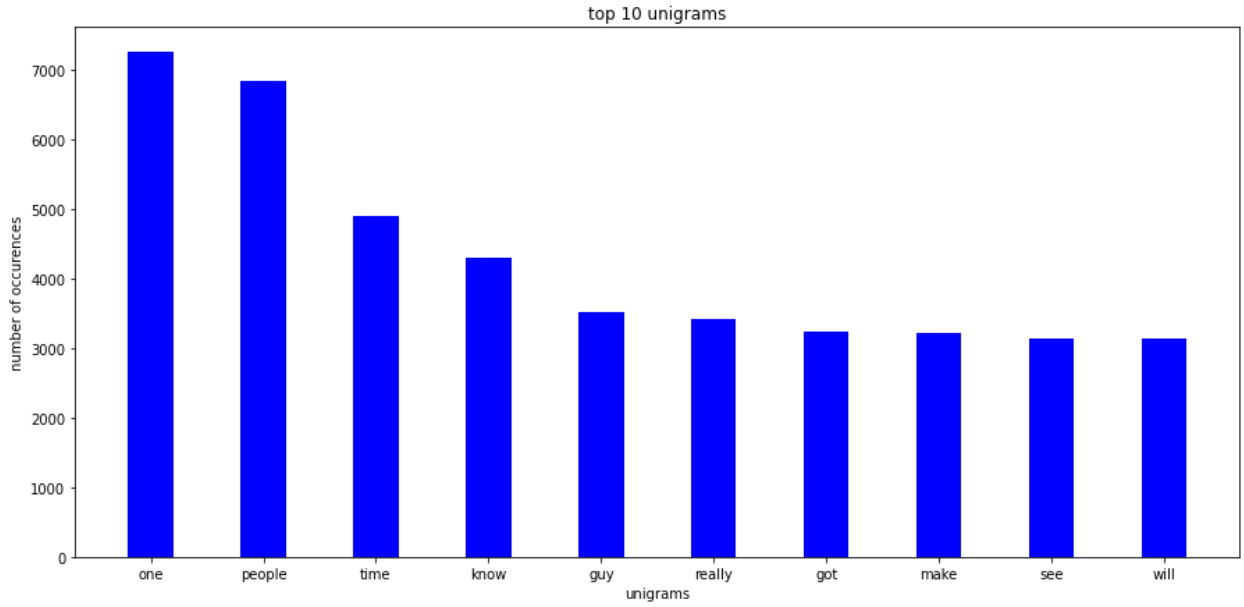
Unigrams for news



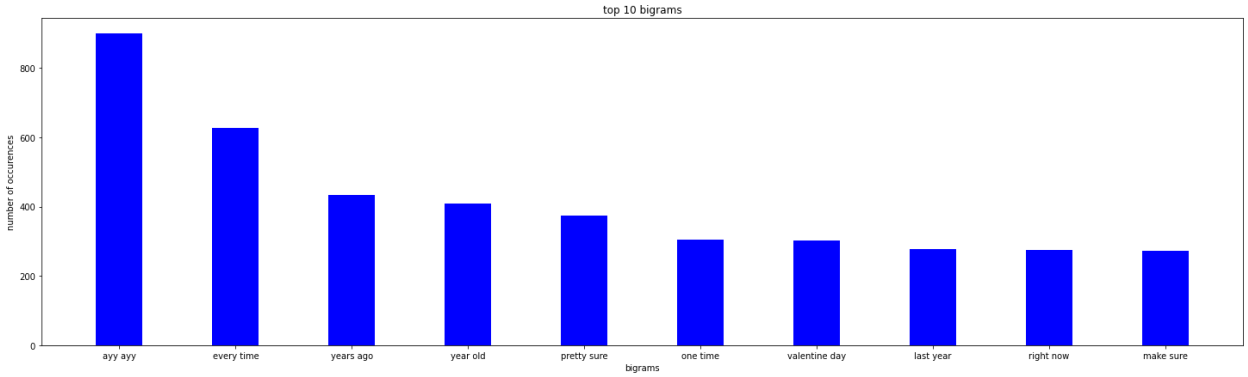
Bigrams for news:



Unigrams for humor:



Bigrams for humor:



Inferences:

1. Unigrams like people indicate the public's opinion in the text corpus.
2. Also, bigrams like Ron paul, Bernie sanders, indicate the presence of political and law disputes between the two parties.
3. Bigrams like hiary clinton and supreme court are due to the us presidential elections.
4. Bigrams like Valentine's day, guy, make, are under the humor subreddit category and indicate funny comments. Also, it can be assumed that these bigrams were used around the month of February
5. Bigram ayy ayy appeared in the humor graph indicating the use of the gram in many humor sentences.
6. Many unigrams can be grouped to infer the context of the text.

Q2

Preprocessing:

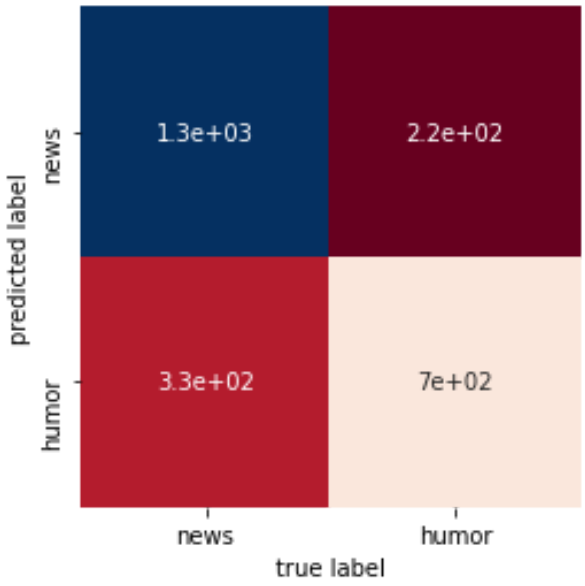
1. dropped all duplicate rows from the dataset: before dropping the duplicate entries, was getting around 99% accuracy so I decided to drop the duplicates to get a more true result on uniques datasets.
2. Added a label column to the dataset which has a value of 0 if the row is under the news category and 1 for the humor category.
3. Split the dataset into train and test data in the ratio of 70:30 using the train_test_split function.
4. Used count vectorizer to implement the bag of words approach which basically turns the sentence into 0s and 1s vector format.

b.

1. Used naive Bayes classifier with the help of MultinomialNB() from the sklearn library.
2. Got the best result for naive Bayes(79% accuracy) so at the end kept it as my final model.

c.

Confusion matrix:



accuracy score: 0.784037558685446
precision score: 0.7625272331154684
recall score: 0.6769825918762089
f1 score: 0.7172131147540984

D.

Correct classification:

1.generally speaking i m an asshole that being said i am profoundly nice to customer service reps whenever i have to call a call center for anything my primary objective is to get them to have an actual conversation with me the reason i called is secondary it ll be handled anyway(humor as humor)

2.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 as humor)

3. i will go out of my way to transfer you to make a second 40 minute car journey to come back for a refund (humor as humor)

4.maybe the bird is a really tough ski coach you call that skiing smack smack beakpoke get up get up and work those poles i do nt actually ski so i m having a hard time coming up with proper sounding ski tips so i ll just keep trying swish those goddamn hips like you mean (humor as humor)

5.this war on cash is such bullshit people carried cash until 12 years ago and it was perfectly fine now they are trying to restrict flying with cash purchasing.. (News as news)

Wrong classification:

Examples:
1.there s really no mystery there were other lines of human beinghomo florensis as they are called as well as neanderthals homo erectus and perhaps the denisovians were all parallel human lines that did nt make it past that last vicious ice age 50000 years ago (news as humor)

2.i had cox for 3 years service was good and over that 3 years doubled my download speeds twice no extra cost upload was still garbage but hey 10015 internet for 65 a month was still pretty killer.(news as humor)
The tone of the sentence is itself lookin as humor.

3damnit i misunderstood the joke at first and tried counting the words he could say per year at first i was like pft no punchline here wordcount is different then i backread fail(news as humor)
Joke is more common to humor

4.it s like a wife seeing her husband with another woman he can swear an oath afterward but the trust is lost klimenko was quoted as saying(humor as news)

5.iwhere in lebanon was this looks a bit like beirut if so i can understand the mixed reactions i would wager the reactions would be less mixed in the northern or southern regions.(news as humor)

Reasoning:

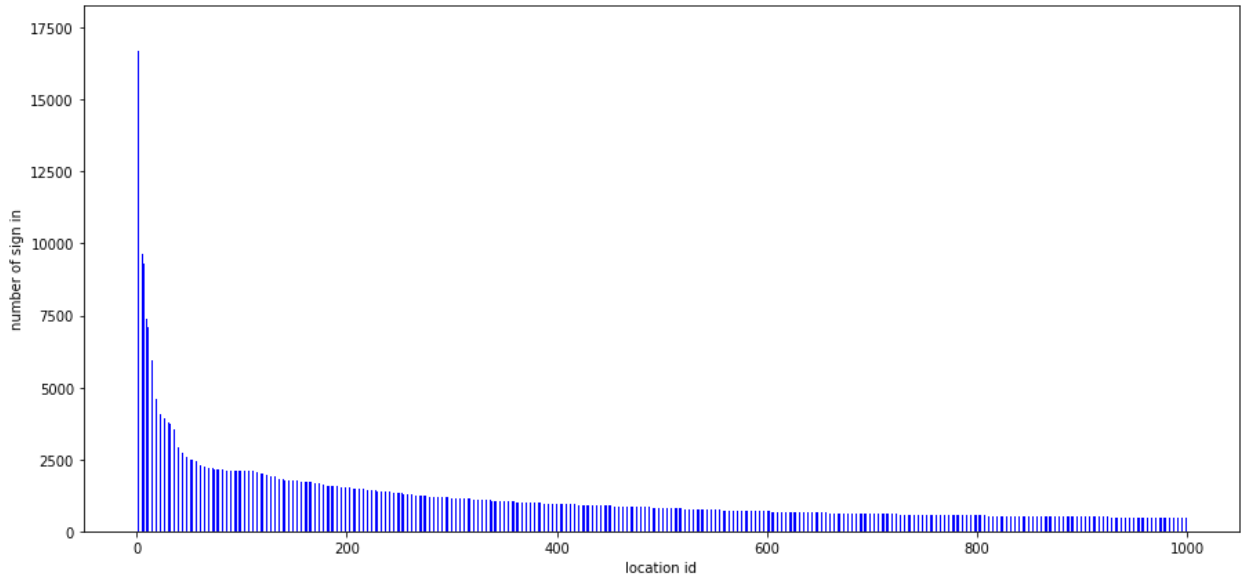
1. In my model, I have used a naive Bayes classifier that operates on probability. If a word appears in a sentence then all the sentences that contain this word are taken into consideration. If this word appears more in the humor category then the sentence will be classified as humor, the same is the case with the news category.
2. Also, it can be the case that the sentence contains popular unigrams from both the categories, in this case, the classification becomes difficult, and hence the model output wrongs predictions.
3. Text that has words that are more common to the news category is classified into the news and similarly for humor. This is not a very good measure as it is leading to false predictions in some cases.
4. Every sentence has been classified wrongly due to the presence of a word that hints the sentence belongs to the other category.

Section 2

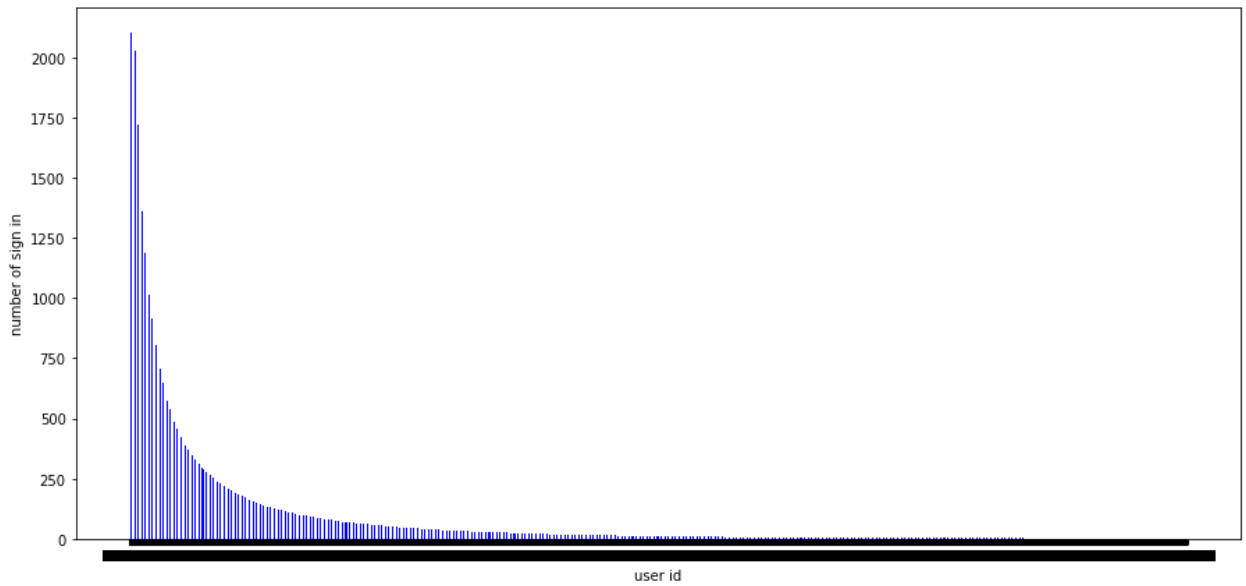
Q1

Removed the locations which had latitude and longitude as 0 because these entries were faulty entries.

a. Plots
    i. Locations v/s No of Check-Ins:



ii. User v/s no of check-ins.

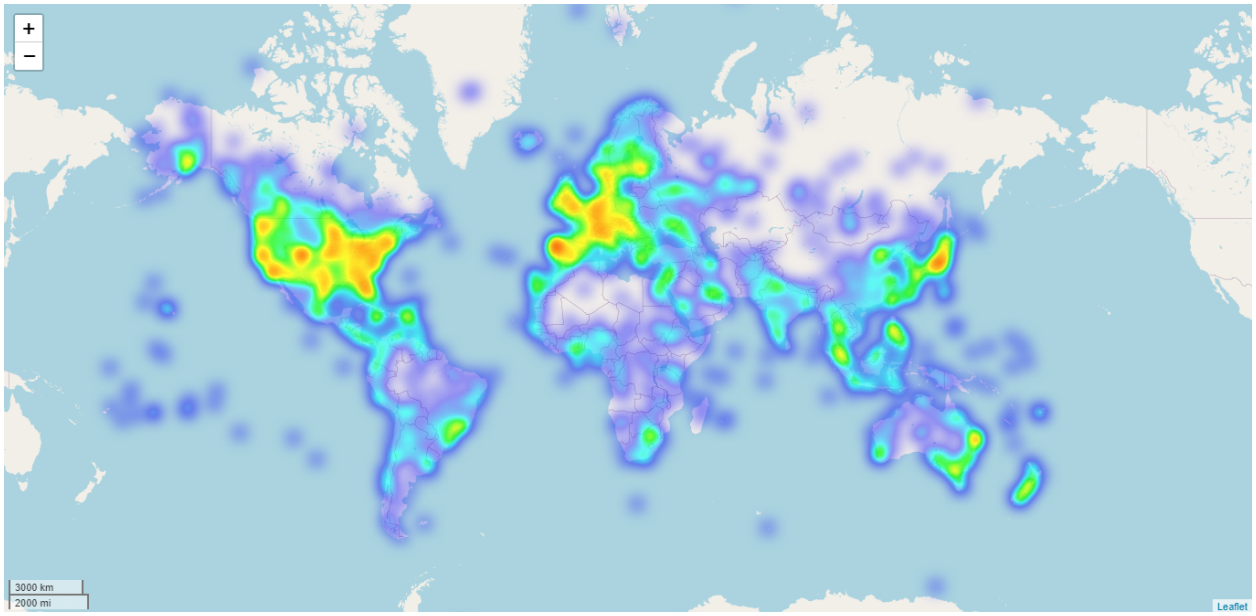


(i have sorted the user-id according to the number of check-ins, the user id is no visible since there are a large number of user id plotted on the graph)

Inferences:

1. Many users had sign-ins around 2100 indicating their usage of the platform is higher than the others.
2. The users that have a high number of check-ins make it easy to track their locations, where they are going, etc. This could be used to track their daily activities and places they visit and hence they can be suggested places depending on their frequent visits.
3. The sites that have a lower number of users visiting them could use this information to track the number of users visiting and take measures to improve their visiting score. For example, restaurants that have lower numbers of visitors could benefit from such data.
4. We have a very high number of locations and different users on this platform suggesting that the platform has been used world widely.
5. The top 5 locations had sign-ins of around 10000 suggesting that these could be places of high importance as many people have visited it like the tourist places.

B. geographical heatmap

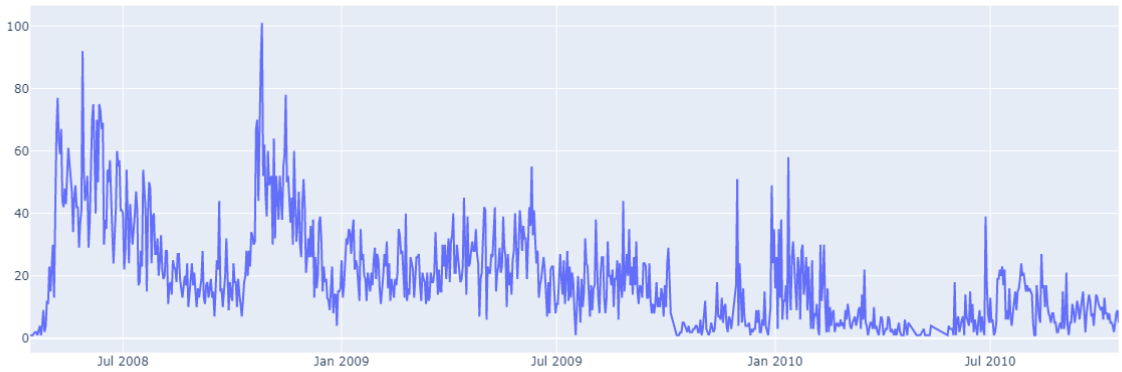


Inferences:

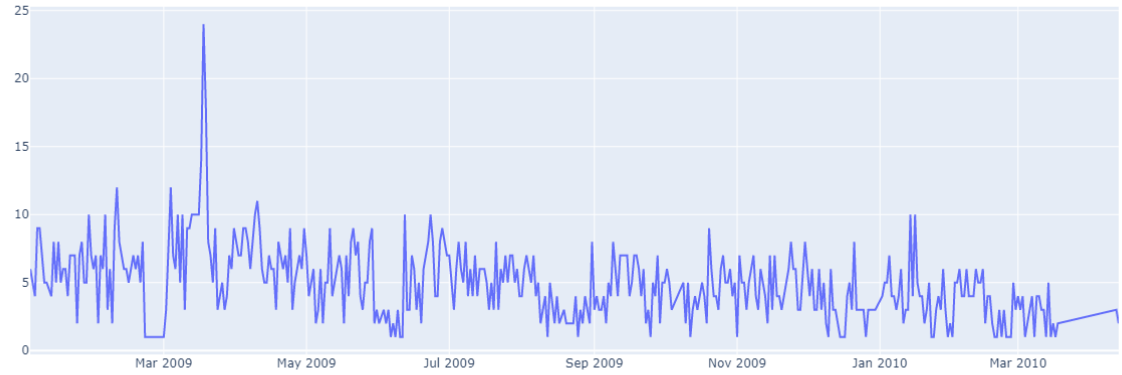
1. As seen from the heatmap that overall every location on the globe has been recorded in the dataset except for parts of Russia, Africa suggesting that places with moderate temperatures are visited more frequently than the places having extreme climates.
2. The most frequently visited places include America, parts of Australia, parts of India, China.
3. If we specifically speak of India the maximum activity can be seen in cities like Delhi, Mumbai, and Bangalore.
4. In America, the most frequently visited cities include, Los Angeles, New York, Washington, etc., and can also be seen in red spots.

c.

i. Check-ins at the location with the most number of check-ins V/S Time
Location id: `ee81ef22a22411ddb5e97f082c799f59`



ii Check-ins of the user with the most number of check-ins V/S Time
User id- 37



Temporal Information that can be used to target users is the following:

1. If correctly analyzed the data can be used to give a personalized experience to the users by suggesting to them the places they might like to visit.
2. This dataset can be used to spread agendas and run campaigns in places that are frequently visited. This way the organizations can reach many people in a short amount of time.
3. This information is also very useful when opening new businesses. Places that have a high number of tourists can be preferred for the location of the business.
4. We can analyze data to see at which period of the month or year the location expects a higher number of visitors. This could again be useful for government and businesses.
5. The peak season for most of the location ids is autumn and summer.

Q2

Leveraging the information:

1. This dataset shows the traveling patterns of the users, which time they travel more frequently, what places they visit more frequently. It also shows what places are more preferred by people and what is the time a location expects a greater number of visitors.
2. All this information is very useful from a business point of view. People can change their restaurants, malls, or any other places according to the places most visited by the users.
3. This information can be used to set up new businesses in areas that have more people visiting them.
4. Also, the business can arrange for extra staff and resources around the time when the place expects the most number of people.
5. This information can also be used to harm people. For example, by observing the traveling patterns of users, robberies and thefts can happen easily.
6. This dataset when analyzed properly can help in suggesting the people the places they should visit which will be suggested by studying the behavioral patterns of the user from the given dataset.

challenges:

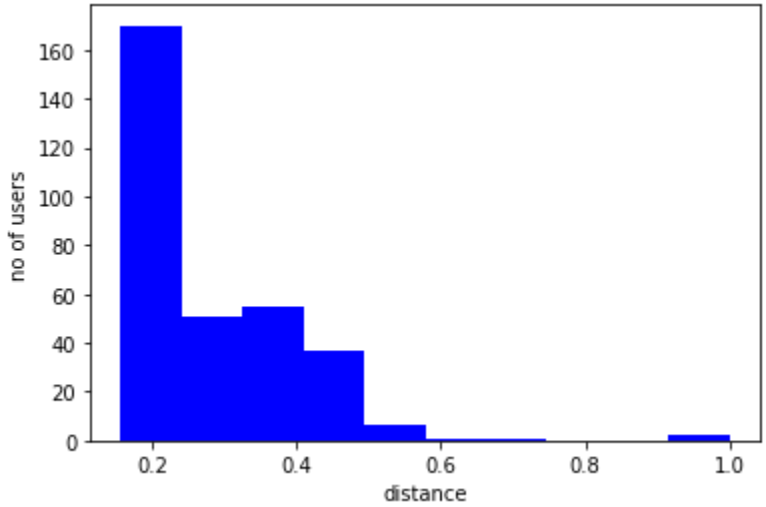
1. Breach of the law: Right to privacy: this information can violate the law as the location the person is in is very private information and it can be used in many ways to harm the user.
2. This data once collected is not restricted to the platform itself, nowadays the platforms sell the user data to other organizations. So extra efforts have to be made to ensure the security of the dataset.
3. These organizations can analyze and suggest to us something in which we are interested. This way we are forced to see many ads on the internet.
4. Knowing the exact location of the user, one can easily steal or even kidnap somebody.
5. Knowing the places that are highly visited, terrorist attacks can be easily planned.
6. In such a platform it is very difficult to gain the trust of the users, so it becomes very important to ensure the security of the dataset, and also the selling of such a dataset can prove to be disastrous.

Section 3

Q1 the metrics used are:

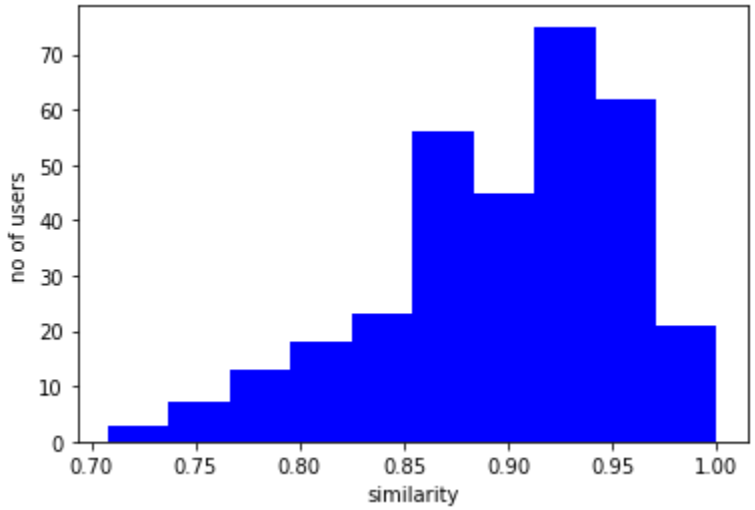
1. Levenshtein distance: it is basically the edit distance between two strings. Ie, the number of characters that do not match between the two strings and in this case, usernames across two social media.
2. Jaro similarity: it returns a float number representing the similarity between the two usernames. Higher the returned value the more similar the two usernames are.

Twitter-facebook
Levenshtein distance:



Mean :0.27372063376434164 Median :0.23529411764705882

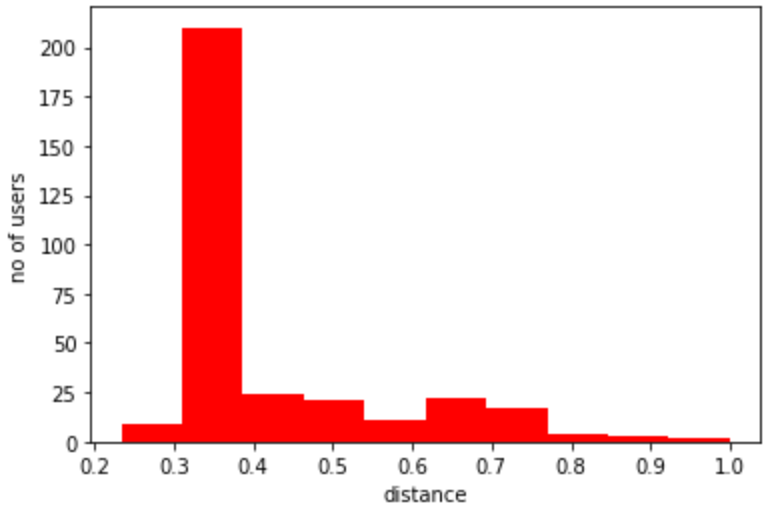
Jaro similarity:



Mean 0.8977813132798821 Median:0.9103483380081844

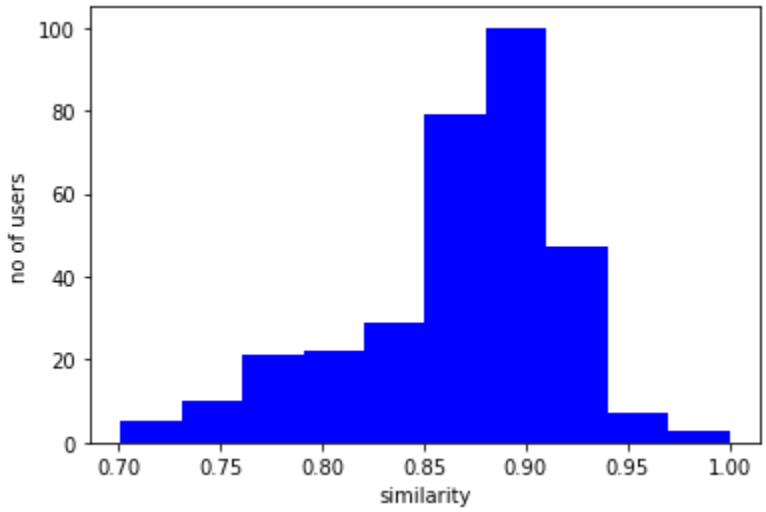
In this case, we say that Jaro similarity is more suitable as the mean is very close to 1.

Twitter-Instagram
Levenshtein distance:



Mean: 0.43219814241486065 Median :0.36666666666666664

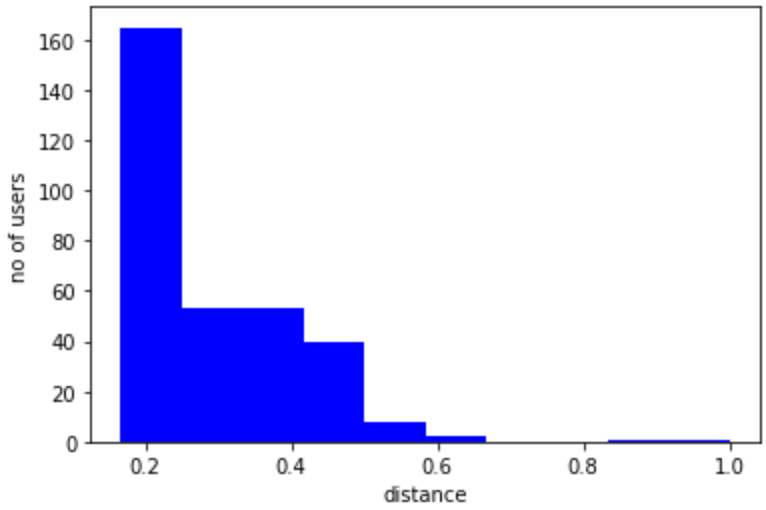
Jaro similarity:



Mean :0.8673978183531125 median: 0.8781021870158012

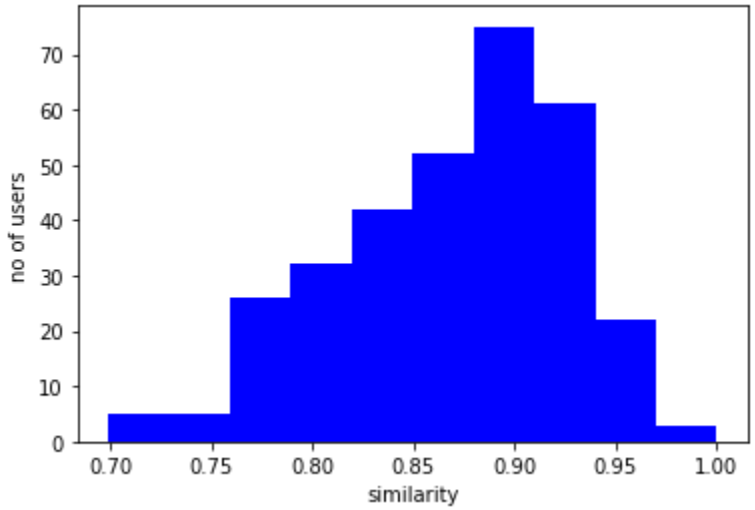
In this case, also Jaro similarity is higher as the mean is closer to 1 than Levenshtein is closer to 0. (Levenshtein should be near 0 to depict similarity and Jaro should be closer to 1 to depict similarity)

Facebook-Instagram
Levenshtein distance:



Mean :0.29721362229102166 median: 0.24074074074073

Jaro similarity:



Mean :0.8687584178530622 median: 0.8784809222716904

In this case, also Jaro similarity is preferred as its mean value is closer to 1 than Levenshtein distance is closer to 0.

Inferences:

1. Jaro similarity gave the best result as they had value closer to 1.
2. The graphs, average, and median values for both the metrics show that most of the usernames match across the different OSM platforms. Hence indicating that username matching across different OSM is a good method for identity resolution.
3. for a user who wants to be anonymous, this could be a point to take into consideration. He should keep his user name different across different OSM platforms. On the contrary, the user who wishes to be recognized should keep his usernames similar.
4. The graphs show more similarity between the pair of twitter-Instagram. This could be due to the fact that on Instagram and Twitter the username is more preferred for searching users and in the case of the name is more preferred.

Section 4

Q1. predicting SSN numbers as compared to the aadhar number is relatively easy due to the mentioned reasons.

1. If we notice the build of an SSN number we can see that it has a certain pattern. The SSN number can be broken down into certain groups of digits.
2. The first 3 sets of digits are based on the area number the person resides in, the next two digits are based on the group numbers and the last set of digits vary from 0001-9999.
3. Knowing the breakdown and pattern in which the SSN number was created helps in the guessing of the SSN number.
4. The SSN number can be predicted easily as the area numbers and the group number are available publicly and hence making the guess of the SSN number easy.
5. Guessing the aadhar number is not easy as 11 out of the 12 digits are generated purely randomly. Hence according to me, aadhar number can not be guessed unless the number itself is compromised.

Q2.

Realizing use principal is difficult from the industry point of view due to the mentioned reasons:

1. Specifically speaking, the harder task is to keep a check on how the data is being used as compared to controlling the access to the data.

2. The organizations that are focused on collecting data and using it in various ways to increase their business collect data that is not even relevant to them. This is due to the fact that this data may come into use in the future. Now since data is stored without any restrictions or any supervising authority, it becomes difficult to implement “use limitation”.

3. Also, these organizations share data between them which makes it even more difficult as the data becomes available to a greater number of organizations.

4. Also due to sophisticated technology the data that an organization stores is shared very securely and hence it becomes difficult to see what data is being shared. This is a pro as well as a con of technology.

5. Also there is a lack of organization of data in the organizations because most of the data that an organization possesses are even not relevant to them.

Q3.

1. This is a fact that we are happy when we get what we want. So by analyzing the likes, shares, comments, and many other things the platform can give us a personalized view. Instead of showing us random things if we are able to see what we like then I think this would not be called a privacy or security breach as the sole purpose of analyzing this data is to benefit the user.

2. Also, the platform gives the user the option to download the stored data at any time. The user then can see whether some irrelevant data is being collected or not. In this way, the platform is transparent to the users.

3. To tackle the arguments regarding sharing of location addresses we can argue that the data is collected to benefit the user as we can monitor the areas and may even warn users of potential threat areas.

4. Also as the provider of the platform it becomes the responsibility of the platform to ensure that users are secure on the platform. In order to do so, it needs information on the basis of which it can classify users as spam and warn other users regarding the same. Also by collecting and analyzing data it can stop the spread of wrong information or even hatred against a particular section of society. Also sharing of pornography can be stopped in this way.

5. Also to keep the platform alive and up it very crucial to give the users what they want. So the minimalistic data the platform collects is important to suggest users the relevant data to them.

6. This is also necessary to the user as suggesting the user with relevant data might get the user some benefits like discounts on retail sites.

References:

https://plotly.com/python/

https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html

https://python-visualization.github.io/folium/quickstart.html

https://scikit-learn.org/stable/modules/naive_bayes.html

https://stackabuse.com/levenshtein-distance-and-text-similarity-in-python/

https://www.geeksforgeeks.org/jaro-and-jaro-winkler-similarity/