Question 1
A) What is a representative dataset?

A representative dataset is a subset of the original dataset which contains more and precise information about the set than any randomly selected dataset. Representative dataset contains selected datapoints from the original set.

Representative dataset is used widely in unsupervised learning problems like clustering.

B) Refer to the definition of representative dataset you wrote above and explain why this dataset does or does not fit into that category. If not, list the challenges in building a representative dataset for this task.

That dataset is a complete data of the user of a particular city. It is a complete set not a representative set. A representative set is a subset of the original data that contains most of the information about the original set with minimum redundancy.

So, it is clear that our dataset is not a representative dataset, so we'll have to convert it into a representative dataset. To convert it into a representative dataset there are various challenges that I found:

1. That dataset contains many things that are not necessary for our representative dataset. For example, it contains various hashtags, links, tags, emojis that are useless. So, for that we need to preprocess our dataset so that it will contain only unique words and all of the redundant and useless things will be removed including stop words also.

2. Our dataset may be skewed either left or right skewed means there may be a trending topic on which everyone is tweeting about in that case our data will not contain much information or uniqueness and thus it will be a problem for us to create a representative dataset.

3. Another challenge was what should be the best size of the representative data such that it will be minimum and covers maximum variance.

4.One more challenge was how should I sample data from my original dataset like what should be included and what should be left out.

These are all the challenges I found out to build Representative dataset for this data.

C) Outline various kinds of biases you observe in this dataset. Substantiate your observations with tables/graphs to effectively visualize the biases.

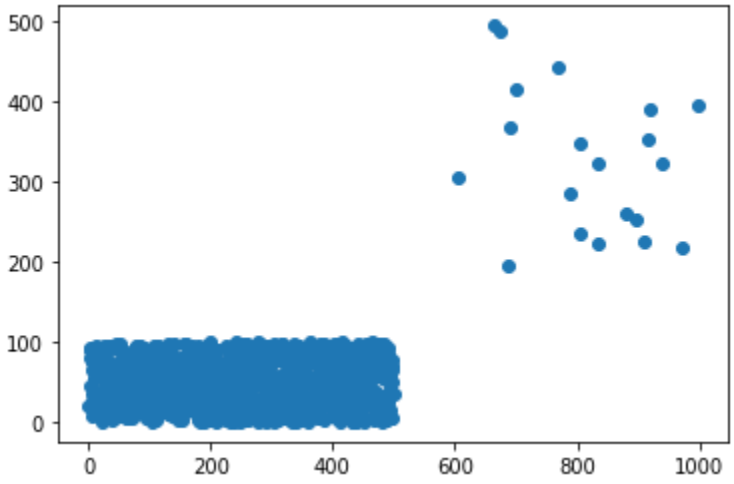
There are various biases I observed in this dataset, some of them are the following:

1. Historical bias: That bias is the most common bias in most of the dataset. Since we are fetching the tweets of the different users it is possible that his/her old tweets will also be fetched so in that way it will be a useless tweet because we were supposed to fetch only the latest tweet so our dataset will contain a lot of historical data which is not relevant.

2. Sampling Bias: Sampling bias means it is possible that some of the dataset is more likely to be selected than others so how we can sample so that any particular sample will not be in the majority.
3. System drift bias: It is quite possible that our system will generate a different set of data each time we fetch it so that kind of bias is called system drift bias.
4. Exclusion bias: This is the bias that occurs when we miss out an important sample from the dataset such that it affects the final representative dataset's accuracy.

These are all the biases that I observed in my dataset.

This is the example of sample bias of my dataset here we can see that a large part of my dataset is of similar type only some of them and more unique and sparser.



D) Propose some methods (at least two) by which these kinds of biases can be avoided while building real-world datasets (for any task).

Some methods to avoid biases:

1. Doing a lot of data preprocessing .
2. Using Different and effective sampling procedures can help us to reduce sampling biases, we can also use a method calls oversampling.
3. We can use random shuffling of the dataset to avoid biases at a certain limit because randomization is the best way in certain areas of analysis and approaches.

Question 2

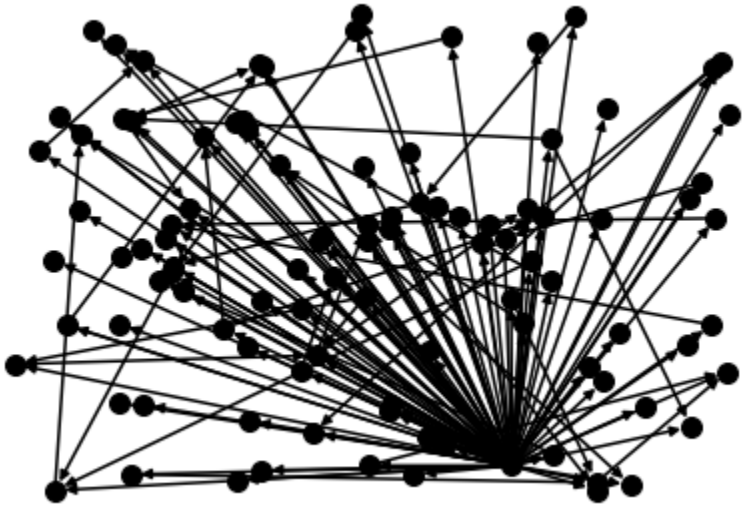
A) How do platforms like Twitter, Facebook and Instagram suggest friends to users? List down a few techniques / methods they may be using.

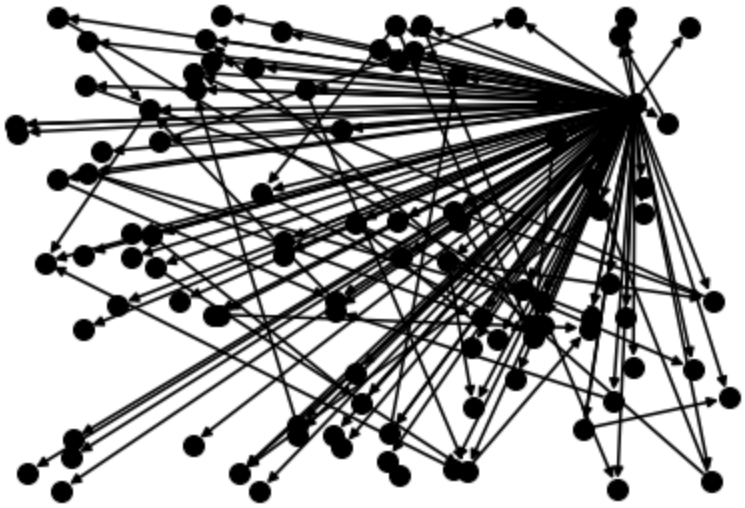
There are various ways by which these social media platforms like Twitter, Facebook and Instagram suggest friends to their users, here are some techniques:

1. By using our contact/ Emails: You never noticed but when you download these apps on your phone these apps ask for the contact/Email Read permissions, and now using this they can read all of the contacts on your contact list and can suggest you those people who are not in your friend list.
2. By using location: These social media platforms using our location can find out what are all the people in the same locality and can suggest these peoples.
3. By Suggesting the people who search for you.
4. By suggesting recently added friend's friend.
5. By Common things between people (School, work, college, city, Tagged in the same picture/post)
6. By network analysis algorithms like they use some kind of BFS (Breadth first search) algorithm. On the first level of the graph, we can find all of the friends of a user, on the second level we can find all the friend's friends and so on.

B)

1. Identify the number of open and closed triads in the graph. Explain any emerging patterns
and differences between these two kinds of triads. You can read this paper to gain some
useful insights about triads and triadic closures.





These are some Graphs I am getting on applying network analysis on 3 different celebrities.

Here we can see no. of closed triads are very less for each of the graphs and no. Of open triads are significantly more than the no. Of closed triads. In the first graph no. Of open triads are 24 and closed triads are 5. In the second graph no. Of closed triads are 7 and open triangles are 20 while in the third graph these no. Are 33 and 4.

Here we can see the pattern that no, of closed tried are always less than no. Of open triads. We are getting less no. Of closed triads because it is very unlikely that followers of a celebrity follow another person who also follow back each other.

2. Write your observations (at least 3) on the nature of the graph (any strongly connected subgroups, commonalities among the nodes, possible biases because of the network structure, echo chambers, etc.)

We can see that these graphs follow the following properties:

1. These graphs are Sparsh Graphs because not everyone in the follow list follows each other.
2. These Graphs are not strongly connected Graphs, but there exists some strongly connected components.
3. We can see there exists one node from which there is an edge to every other node that node is our celebrity node.
4. If we remove all edge other than celebrity nodes edges then we'll get a star graph.
5. There exist some nodes in our graphs from which there is no outgoing edge means they do not follow anyone from that celebrity's follow list, and they do not follow our celebrity also.

Question 3

Keeping Sahana's Extreme / Hate Speech lecture in mind, please review this paper and list down

the following:

A) At least 2 limitations / challenges (if you find more, please feel free to list) in the conclusions that the paper makes, i.e., paper is claiming more than what it really is.

In this paper they are claiming that "Islam has been studied as terrorism in the past." but that is not completely true. That is somewhat true that there exist cases where people often consider that if a person belongs to the ISLAM, then he must be a terrorist but that is not the thinking of a large number of people. Recent studies found that people actually changing their thoughts about Islam, in this context a small group of people also conducted one social science experiment on what people think about it they did an experiment that was related to the same work that is done in the paper and found that the result was quite positive that people's thinking is actually changing about Islam and this community.

And the second thing they mentioned that "religion was used as a symbol of patriotism" that is also I can say is not completely true because they divvied the data on a random basis and now that data may contains different tweets from different people belonging to different set of beliefs. First of all, there are various bias possible for that sampling such as sampling bias or Exclusion or maybe that dataset can be skewed also.

B) At least 2 ways by which you can address the limitations you mentioned about; be precise and explain the data / methodology changes that you will make to fix the limitations.

By using this methodology limitations can be fixed:

They should consider that they are using a representative dataset for this work, because there are various biases that can affect the result so they should follow the random sampling.

Another thing is they are collecting the data of 224,229 unique users consisting of 410,990 tweets have an average of 34 words in each tweet. Now how they are selecting these 224,229 users is not mentioned there is also a possibility that here also most of the users may belongs to same set of beliefs or perceptions, so there should be a good way to select the users so that there can be different class of beliefs or ideologies that one follows and so that we can get better results.

Question 4: Telegram groups are a common messaging platform and are used as discussion forums for various topics. Multiple groups are created and content/links/images etc. are

extensively shared/forwarded on a daily basis. Consider that we wish to collect data pertaining to the spread of misinformation during Covid-19 on telegram.

1. What methodology can be used to exhaustively collect all data from such telegram groups for this? Discuss the details of the same.

Telegram is a messaging service that allows users to send/receive messages as well as to create groups/Channels where many people can collectively receive messages. This is a very good platform, but it is nowadays used for spreading false information among people. To collect such data, we can use various methods. One such method is we can list out the groups/channels related to the covid information, and we can use one of the telegram Api available to fetch the data of these groups/channels and we can store that data. Now we can train a machine learning classifier which takes input one such post of the group/channel and outputs whether it is covid related or not. To do this we need a predefined dictionary of words containing words that are most likely to occur in a covid related message. After this. After this we'll train this classifier. This can be done for images also. Using this method, we can collect data in bulk and can check whether it is fake or not by monitoring it or by manually verifying it.

2. Mention 2 other use cases where this methodology can be employed.

This method can be used to collect data of various things, for example:

a) To collect data of latest movie release.
b) To collect data for Piracy detection
c) To search for study materials available on different telegram Group/Channels.