Student Name: Chaitanya Kadali, Sai Ram Naik
Student ID: 2019CS50435, 2019CS10340

Database Management (COL362)
Milestone-1

The project we are doing is almost similar to twitter application. Our application project "Freak" also has users and they can follow other users, or tweet and also report tweets by other users. We also have an option of liking tweets by other person. Our application consists of 4 main entity sets i.e Users, Tweets, Links and Hashtags.
The following sections contain ER Diagram, And an overview of the modified database we created from the database we used.

ER Diagram

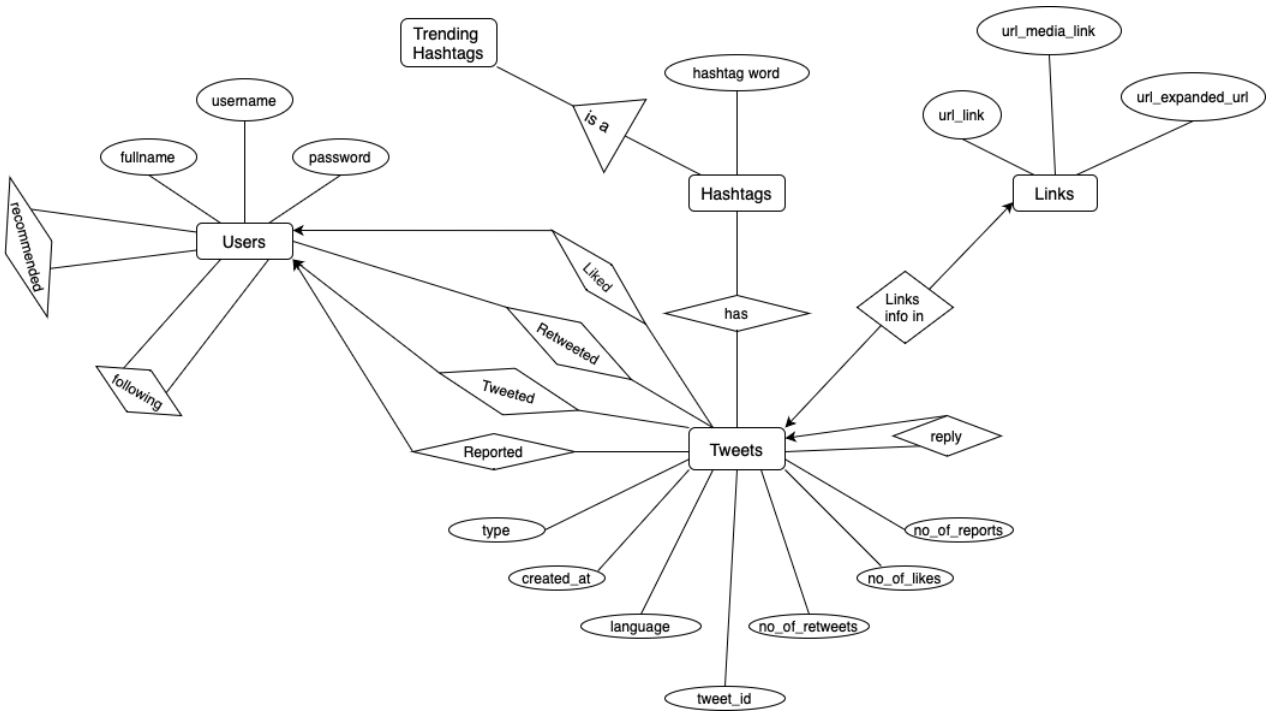

Figure 1: ER Diagram

Main Tables in our Database

The data for this project is taken from https://data.mendeley.com/datasets/7ph4nx8hnc/1
Eventhough this data is collected during the pandemic period and the tweets are related to covid pandemic, this still contains the users personal details and also large enough to create a database social network application. Also we would like to add login and create account options, and this further increases our database size.

Entity sets	Attributes
Tweets	Tweet_id Type Created_at No_of_likes No_of_retweets No_of_reports Language
Users	fullname username password
Links	Tweet_url (link of the tweet) Media_url (Link of the media in the tweet if exists) url_expanded_url (url in the tweet if exists)
Hashtags	Hashtag word

Refining

The original raw data from the link given above consists of 8 tables. And we have extracted 6 tables for our project purpose. Those tables are listed below.

1. **Users** : contains all the personal details like username, fullname, password of a person. This information is extracted from Tw.nodes.csv
2. **Tweets** : contains the information about the tweets. This information is extracted by Joining User.And.Type.csv and Tw.Date.Lang.csv
3. **Links** : contains all the information about the links. This information is extracted from Links.Media.Tweets.csv
4. **Hashtags** : contains all the information about the Hashtags. This information is extracted from Hashtags.csv
5. **Followers** : contains all the information about the followers of a person. This information is extracted by self joining of Tw.Edges.csv
6. **Edges** : extracted from Edges.csv

Table name	No.of Rows	Size before Refining	Size after Refining
Users	3565514	3565514	3565514
Tweets	8982694	8982694	8982694
Links	8982694	8982694	8982694
Hashtags	3239024	3239024	3239024
Followers	921622	921622	021622
Edges	921622	921622	921622