CSE519 - Project - How Much Do People Sleep ?

***Abstract*—The purpose of this project is to determine the US demographics sleeping pattern using twitter activity of users. Globally, every second on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year [1]. Analyzing such huge quantities of data, to predict the sleep pattern would incur high computational overhead on systems. To do a meaningful analysis with limited resources we have used a smaller subset (1%) [2] of global twitter data and have clustered the data based on different demographics such as cities, time-zones. The data classification is thoroughly done based on the tweet's time and location by filtering tweets from bots and managed accounts (considering this would lead to wrong results). From classified data we have found the sleeping patterns for different time zones and cities in US.**

I. Introduction

Social media use is receiving increasing attention independent from other forms of electronic media use, so using this data we can understand/derive the sleeping pattern associated with the individual [3]. With more than 330 million global monthly active users, Twitter is one of the biggest social networks worldwide [4].

From media personalities to politicians, public users to bots, the social network has become a go-to for real-time information and reactions to current events [5]. As twitter is ubiquitous this study tries to predict on how much people sleep based on the demographic tweet activity. Firstly, social media (and other technology use) may directly displace sleep by delaying bedtimes, resulting in shorter sleep duration [6].

Furthermore, associations between social media use and sleep are complex, likely involving interactive and bidirectional effects, for example including the use of social media as a sleep aid. Also, as per Pew research center analysis only around 22 percent of American adults today use Twitter, and they are representative of the broader population in some ways. So, we can use the twitter activity as a large-scale source of data to effectively extrapolate on how a city/demography sleep. There have been numerous studies which uses twitter data to study the sentiment analysis, also there has been a study between late-night tweeting and the next day game performance among basketball players [6]. There have been comparatively fewer studies on sleep analysis vs twitter activity. In this project we have tried to mainly analyze the sleeping pattern of US demography.

We have considered few US cities based on the twitter activity. A graphical plot of twitter activity in the months of Aug, Sep, Dec – 2015 and Jan -2016 is shown in 1. From the figure we can clearly see that the twitter activity in California (PST in general) and New York (EST) is more compared to others. Also, these months were considered as we wanted to analyze if there is any difference in the sleeping pattern between fall and winter seasons. Also, by choosing these months we can get a clear insight on how people sleep during the holiday season.

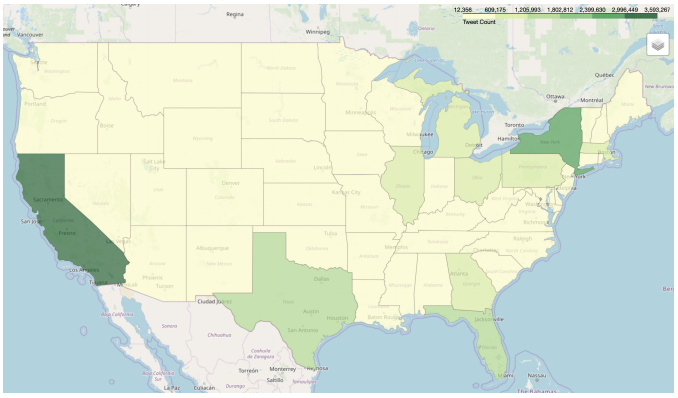

Fig. 1: US Tweets Map

Based on the tweet activity as shown in the graph we identified US cities and time-zones where the activity is maximum and hence have performed our sleep analysis on these cities. As per the data available we have done our analysis on cities like Los Angeles, Chicago, Las Vegas, Houston, Miami, Phoenix & Atlanta also we have considered 4 time-zones – EST, CST, MST and PST which span across the US. The paper covers different section of analysis. Initially, we talk about the dataset, cleaning of the data, data pre-processing and data wrangling. Data cleaning and pre-processing was one the most important part of this project as this ensures that the data gets clustered accurately as per the requirement. As twitter data can also have bots involved and other imperfections like the location anonymity, which leads to wrong analysis we had to do a thorough job during data pre-processing.

The next section of the paper talks about the results obtained. The section clearly highlights the important findings about the sleep pattern across USA. A passel of interesting plots is explained in this section. We also have verified few of our results from the online available data which the paper talks about in the conclusion section.

Lastly, the paper talks about future work that can be extended based on our analysis. We have considered only a small subset of tweets for USA. The same idea can be extended to very large twitter data sets. Also, sleep plays an important role in general health. Studies can be done on general public health by analyzing how the population sleeps.

II. DATA MUNGING AND DATA PROCESSING

As already discussed we have used the data set provided by Prof. Jason Jones, for the months of Aug, Sep, Dec-2015 and Jan-2016. Additionally, we also have scrapped the sunrise and sunset data for corresponding months by using beautiful-soup from timeanddate.com [7]

As part of data pre-processing, the steps involved at a high level are as follows,

- Filter the data based on US location.
- Drop unneeded, redundant and deprecated columns.
- Pass this through a custom bot detector and filter the tweets which are possibly from a bot.
- Filter the tweets which has location anonymity.

Using the above cleaned data-set as a base, doing feature engineering we have added few important columns like timezone, local_time, week of the month. On this data-set the tweet data is clustered based on different time-

zones and also based on different cities.

A. Filter tweets based on US location

The twitter data which is avilable was in the form of a json object. We have used rapid-json which is an extremely fast C++ JSON parser and serialization library [8] to validate and parse json data. After the json is validated against the schema the filtering algorithm is called. A high level overview of how the filtering is done is shown in the algorithm 1. In the twitter data, the location column doesn't have any standard format of input. For example the user can enter any of these strings for the location LA - USA, LA or Los Angles, CA or California, LA. We have split the location column based on the Tokenizer and have compared with a detailed list of US cities and states. If we get a match then that tweet belongs to a US demography. The time-zone field in the twitter data is null for most of the records and so we cannot directly use this field to determine the tweet location.

Algorithm 1: Filter tweet data based on location and bots

```
us_tweets = []
for each line in tweet.json do
    if line starts with "created_at" then
        if check_US(line['location']) &&
          ! check_Bot(line['id']) then
            us_tweets ← line
        end
    end
end
```

B. Filter the data on potential bots

The above obtained data-set is now passed through a custom bot detector which detects whether the given twitter-id is a potential bot or not. We have used the botometer python api [9], which gives a universal bot score to each twitter id considering many parameters like followers count, friends count, time of the tweet every day etc., We have manually validated the accuracy of botometer api by testing it with couple of bot-id's. From this analysis we concluded that, for any id if the universal score is ≥ 0.15 it can be considered as a potential bot and hence we are filtering out these records.

Algorithm 2: check_US and check_Bot

```
us_state_list # list of all states of us
us_city_list # list of all cities of us
check_US(x) :
if x is in us_state_list or us_city_list then
    return True
end
check_Bot(x) :
result = botometer.check(x)
bot_score = result['scores'].get('universal') if
 bot_score ≥ 0.15 then
    return True
else
    return False
end
```

C. Clustering and Feature Engineering

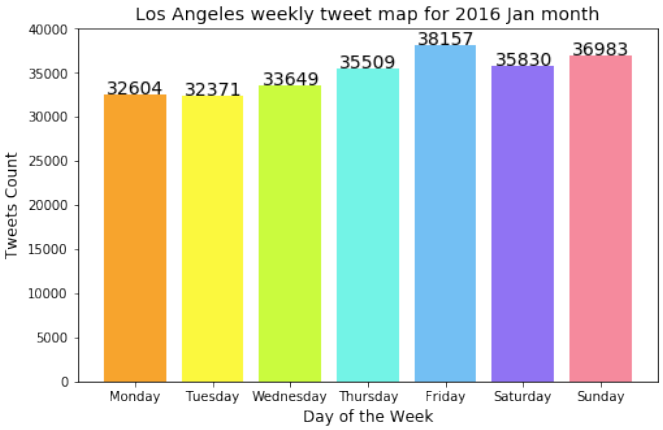

Fig. 2: LA : Weekly tweet trends Jan'16

Now the next task at hand is to properly categorize data as required. The data-set is now of only US based demographic and is free of bots as this has been filtered out. Based on the us location available we have

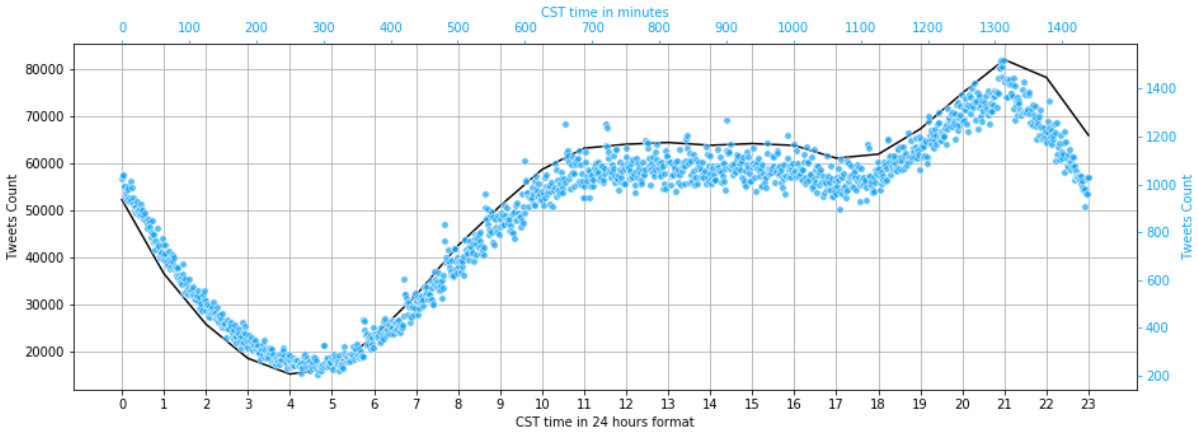

Fig. 3: CST Tweet Pattern - Aug'15

added the time-zone which corresponds to that location. Also, there was some more data cleaning required here as some of the tweet record's location was just set to US/USA. Based on this we cannot determine the timezone, so all these records had to be dropped.

Now, we have the tweet created_at time and the corresponding time-zone information for all the us tweets. Using this information few more important columns were added like "local_time" which corresponds to the local_time the tweet was made, "day" of the month, "hour" , "min" and "week" of the month. Using these we have classified the twitter data into 4 time-zones (EST,CST,MST,PST). The Alaskan time-zone was not considered as number of tweets from this time-zone in the original 1% data sample was too low. Therefore, the sleep analysis has been exhaustively performed on these 4 time-zones.

Next in order, few major cities from these time-zones are considered (Houston, Los Angeles, Chicago, Atlanta, Las Vegas, Miami, Phoenix). Sleep analysis on how much these time-zone's / cities sleep in a month, on a weekday or weekend is explained in the next section.

In this section we can look at 2 clustering examples, Figure 2 describes tweeting trends w.r.t to city and figure 3 shows a tweet pattern for CST time-zone. From the LA city tweet trends graph we can see that the number of tweets during weekends increases when compared to weekday tweets. The sleep analysis of LA revealed that, on weekends people tend to sleep later in the night than usual.

From the twitter activity for CST time-zone we can see from the graph that the tweet activity decreases (below 15%) around 12.30am. In the results section, we talk on how to determine the sleeping range.

III. Results

The algorithm 3 explains on how to determine the sleep range for any demography given the tweets per minute. We consider the general population of a demography sleeps when the tweet count goes 1 standard deviation below the mean. This can be considered as a reference point below which people start to sleep. The algorithm extracts the time of the

Algorithm 3: Sleeping range in minutes

```
val = tweets.mean - tweets.std;
SleepStartMin = -1;
SleepEndMin = -1;
for All day mins do
    if tweets[index(mins)] ≤ val then
        MinsList.append(mins);
    end
end
Sort(MinsList);
if MinsList contains consecutive 15 mins then
    SleepStartMin = first min of consecutive 15 mins;
end
Sort.Reverse(MinsList);
if MinsList contains consecutive 15 mins then
    SleepEndMin = first min of consecutive 15 mins;
end
if SleepStartMin < 0 or SleepEndMin < 0: then
    print("Can't determine sleeping pattern based on
     given data");
else
    SleepRange = SleepEndMin-SleepStartMin;
end
if SleepRange < 0 then
    SleepRange += 1440
end
return SleepRange;
```

day in minutes where the tweet count has fallen below the reference point. Also, the algorithm considers a windows of 15 min of low tweet activity starting from the reference point to safely judge that the demography is asleep. The same behaviour is considered while waking up. We have considered the window to be 15min as normally the sleep latency for people to fall asleep is around 10-20min [10].

A. Sleep-Range

So the algorithm 3 takes tweet data-frame as an input and return sleep start time, sleep end time and the sleep range. Using these values detailed sleeping pattern analysis has been done for different time-zones ans also for different cities.

The analysis includes, sleep range over a month, sleep range during weekday v/s weekend. Also, few interesting sleep analysis for different weeks for December month has been carried out.

B. Time-zone Analysis

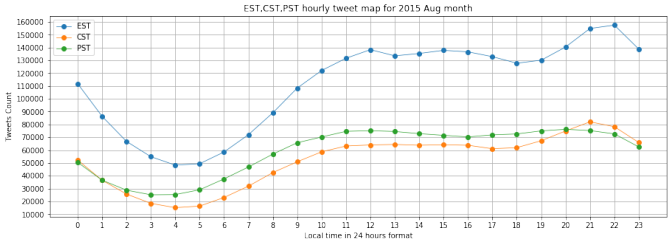

Fig. 4: Time-zone Twitter Activity - Aug'15

The plot in figure 4 shows the tweet trends for the month of august across time-zones. As seen from the graph there is a huge dip for all time-zones during the early hours of the day. From this we can infer that the demography is asleep around that time. The twitter activity gradually increases as the day proceeds and the twitter activity spikes during late evening (After 8pm). This trend accounts for the fact that, people use twitter a lot more after normal working hours. As people fall asleep during night, the twitter activity goes down as seen from the graph.

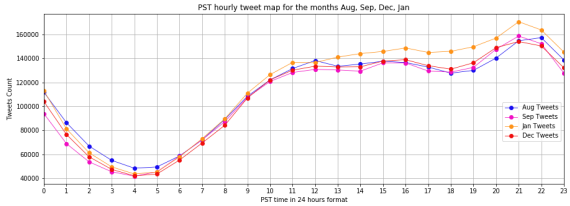

Fig. 5: PST Sleep Plot for all months

The figure 5 which is for the time zone PST show that the twitter activity across different months follow the same trend. Though the

trend is same we can see that number of tweets has increased from Aug'15 to Jan'16 this might be due to the addition of new twitter users during the time period Aug-Jan.

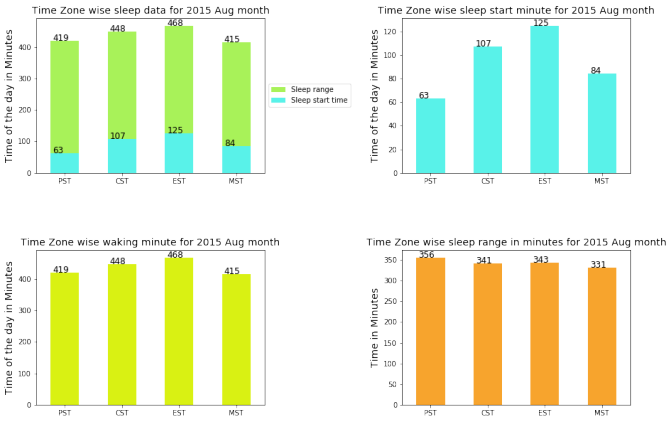

Fig. 6: Time-Zone wise sleeping pattern - Aug'15

In the figure 6 we have plotted the sleeping start time, sleeping end time and sleeping range in minutes of the day. The sleep-range algorithm 3 used to get the sleep range for all time zones and corresponding values have been plotted. From the plot we can see that majority of the population in PST time-zone are asleep around 63rd min of the day (which is 1.03 AM), while the population belonging to EST time-zone sleep around 125th min (2.05 am). We can clearly see a hour difference in sleeping time between PST and EST time-zone. Even though we see a difference in the time when people go to sleep, the total duration of sleep (sleep range) for both the time-zones are similar.

We have listed down the sleep pattern for all the time-zones, for the months Sep, Dec-15 and Jan-16

September month sleep analysis

Sleep Parameter/Timezone	PST	EST	CST	MST
Sleep Start	39	76	70	51
Sleep End	389	447	418	406
Sleep Range	350	371	348	355

December month sleep analysis

Sleep Parameter/Timezone	PST	EST	CST	MST
Sleep Start	41	98	99	77
Sleep End	411	462	449	419
Sleep Range	370	364	350	342

Jan month sleep analysis

Sleep Parameter/Timezone	PST	EST	CST	MST
Sleep Start	51	104	100	68
Sleep End	403	463	453	418
Sleep Range	352	359	353	350

All the values in the above sleep analysis tables are in minutes of the day (00 - 12am). From the above tabular analysis we can see that, even though people sleep during different times for different months (this may be due to sunlight and seasonal effect), the overall sleep range is almost same.

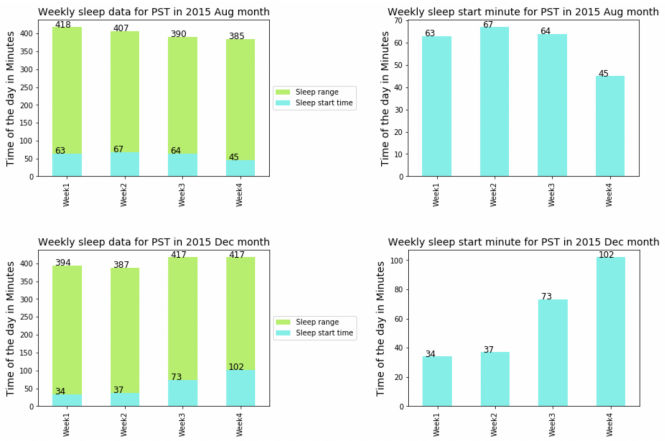

Fig. 7: PST Weekly Sleep plot - Dec v/s Aug

We analysed the weekly sleeping patterns for different time-zones. We found an interesting insight for weekly sleeping pattern of December month. The figure 7 shows a weekly sleeping pattern comparison

plot for December and August month of PST time-zone. We can clearly visualize that, people sleep late at night for the last two weeks of December when compared with august data. One of the major factor influencing this sleeping pattern change could be from the season holidays in December month.

C. City wise Analysis

In this section we talk about the city wise sleeping pattern analysis. In the same lines as

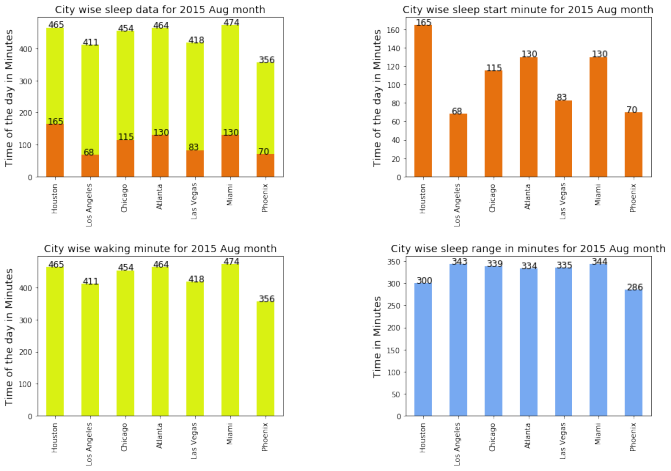

Fig. 8: City wise sleeping pattern - Aug'15

above, we continued our analysis to predict the sleep pattern for different cities of USA. The graph in figure 8 shows the sleeping pattern for different cities for the month of Aug-15. We can see that LA sleeps earliest around 1.08AM which Huston sleeps around 2.45AM. We can see a strong correlation of sleeping pattern with the cities respective time-zones. From our analysis we inferred that Phoenix city sleeps the least when compared to other cities. The cities LA, Chicago, Atlanta have similar sleep range.

We have listed down the sleep pattern for all the cities, for the months Sep, Dec-15 and Jan-16

September month sleep analysis

City/Sleep Parameter	Sleep Start	Sleep End	Sleep Range
Houston	95	418	323
LA	57	394	337
Chicago	81	418	337
Atlanta	107	468	361
Las Vegas	48	386	338
Miami	129	419	290
Phoenix	41	358	317

December month sleep analysis

City/Sleep Parameter	Sleep Start	Sleep End	Sleep Range
Houston	138	477	339
LA	63	409	346
Chicago	110	433	323
Atlanta	131	476	345
Las Vegas	68	374	306
Miami	111	447	336
Phoenix	53	416	363

Jan month sleep analysis

City/Sleep Parameter	Sleep Start	Sleep End	Sleep Range
Houston	134	449	315
LA	50	412	362
Chicago	92	435	343
Atlanta	113	478	365
Las Vegas	64	397	333
Miami	112	471	359
Phoenix	91	409	318

From the above table we can clearly see that people start sleeping late at night in all cities during the months of December and Jan when compared to Sep. As seen, the seasonal effects are more prominently visible in city wise analysis than the time-zone wise analysis. The below tabular column plots the season changes w.r.t sleeping pattern of Houston city.

	September	December
Sleep Start	95	138
Sleep End	418	477
Sleep Range	323	339
Avg Temp(F)	80	61
Sun Set	7.20PM	5.25PM

Though the temperature drops during December month and the Sun sets early people in Houston tend to be awake late in the night. This might be due to the fact that December month gets more holidays than September, and people tend to tweet more on holidays/festivals.

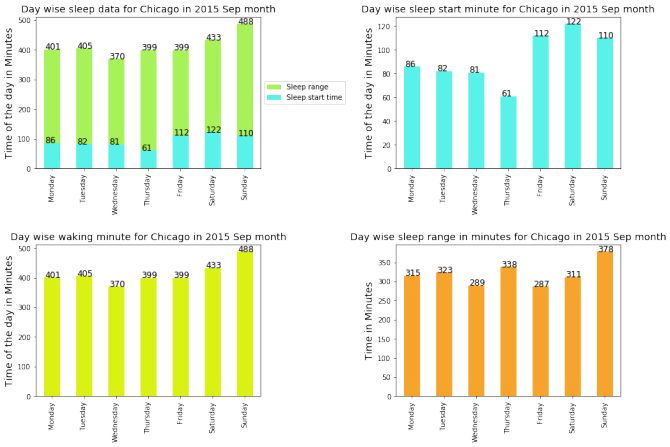

Fig. 9: Chicago daily sleeping pattern - Sep'15

Fig. 10: Sleeping pattern for weekday v/s weekend - Fitbit analysis

We did a sleep analysis based on days of the week for Chicago city. As seen from the figure 9 people tend be awake late in the night during weekends when compared to weekdays. One can see that people tend to sleep (sleep-range) more on Sundays than other days. We compared our analysis with the sleep-jetlag data provided by fitbit analysis [13]. Our analysis of sleep time shift between weekend and weekday correlates well with the analysis performed by fibit which can be verified both figures 9 and 10.

IV. Conclusion

In this report, we have successfully analyzed the sleeping pattern for different demographics of USA. We have given multiple examples on sleeping pattern using twitter data which was clustered based on time-zone and also on different cities.
The analysis put out in this paper highly correlates with the sleeping pattern stated out by Jawbone sleep tracker [11] [12] as shown in the figure 11

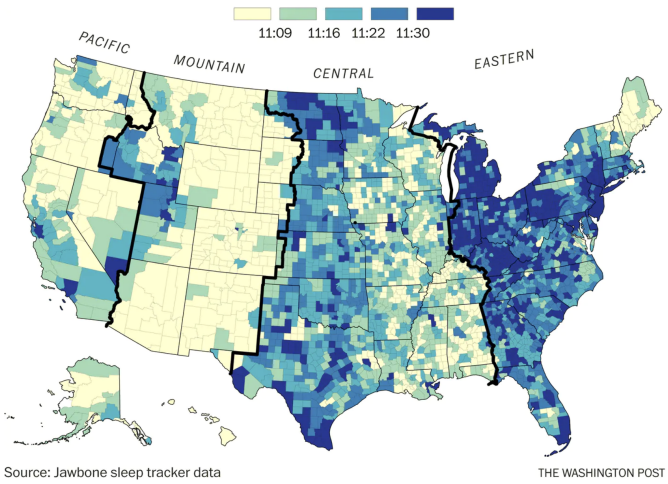

Fig. 11: Time-zone sleeping Pattern by JawBone

The time-zone wise sleeping pattern analysis by us and the JawBone sleeping data shows similar sleeping trends. We can observe that from both the data, PST sleeps earliest followed by MST, CST and EST.

V. Future Work

In future this project can be extended to predict sleeping pattern of the global population by working on very large data-sets. Also, a strong model can be constructed which predicts the sleep pattern of a specific area by using external factors which contribute to sleep like sunset, amount of sunlight received, average temperature during night. Moreover, the sleeping pattern of a demography can tell us about the general health of the people in the region. So sleeping pattern derived using Twitter offer exciting potential for their use in studying and identifying both diseases and social phenomenon.

References

[1] https://www.dsayce.com/social-media/tweets-day/

[2] Prof. Jason's 1% Twitter Data-set

[3] https://www.statista.com/statistics/274564/monthly-active-twitter-users-in-the-united-states/

[4] https://www.odwyerpr.com/story/public/13117/2019-09-23/deep-dive-into-us-twitter-users.html

[5] https://www.sciencedirect.com/science/article/pii/S2352721819301457#bb0025

[6] The Use of Media as a Sleep Aid in Adults - Exelmans, L., Van den Bulck, J

[7] https://www.timeanddate.com/sun/usa/new-york?month=12&year=2015

[8] https://github.com/python-rapidjson/python-rapidjson

[9] https://github.com/IUNetSci/botometer-python

[10] https://www.sleep.org/articles/how-long-to-fall-asleep/

[11] https://sleeptrackers.io/jawbone/

[12] https://www.washingtonpost.com/business/2019/04/19/how-living-wrong-side-time-zone-can-be-hazardous-your-health/

[13] https://www.weforum.org/agenda/2018/02/fitbit-analyzed-data-on-6-billion-nights-of-sleep-with-fascinating-results/