Analyzing Circadian Rhythms through Social Media Activity

A Twitter- Based Global Study

By Lazaro Martull

Introduction

1.1 Project Introduction

Recently social media networks have evolved into instruments for understanding behavior due to their ability to provide vast datasets on user activity trends. Through this project we delve into rhythms which are our natural clocks controlling sleep patterns and body functions by studying Twitter usage as an indicator of global sleep and activity trends. Our goal is to uncover how variables like time zones, job schedules and various demographics influence rhythms in communities. With media late at night comes the risk of disrupting our natural sleep patterns and negatively affecting our health and daily performance.

Twitter is a platform with a global reach that provides valuable data on various disruptions and behavioral trends for analysis purposes. Our study is centered on uncovering rhythm trends through Twitter analysis to gain insights into how diverse demographic groups engage with the platform. Sectors such as transportation and healthcare could leverage this information to improve safety measures and overall employee health by understanding rhythms. This project aims to highlight the rhythms of demographic groups and offer useful insights for applications related to health and productivity.

1.2 How to Use This Report

This report has two objectives; Firstly it offers a summary of the design and execution of the Twitter based study aimed at examining circadian rhythms among various demographic groups;

secondly it presents a methodological model, for utilizing social media data to investigate sleep cycles and other biologically influenced behaviors in everyday life scenarios rather, than controlled environments typically seen in traditional circadian research.

1.3 Essential Background Information, for Novices

To grasp the content of this report effectively it is important for readers to have a background in data science in areas like data preprocessing time series analysis and visualization methods. Knowledge of Python and tools like Pandas, Matplotlib and Seaborn can aid in understanding how data is manipulated and presented. Additionally, understanding rhythms and their influence on human wellbeing can provide context for the findings presented in the report.

The report is structured in this way; The second part explains how data from Twitter was collected and processed using methods such as timestamps and extracting user activity patterns and demographic categories. The third part visually presents trends in Twitter activity to explore how daily rhythms vary among demographic groups and locations. Finally, the report discusses the findings' significance, lays out limitations and suggests uses for these insights, in industries and public health.

Implementation

This research employs an approach that includes gathering and examining Twitter data to explore variations in rhythms and sleep habits among different demographic segments effectively. To meet our goals we have integrated strategies for managing data conducting analysis and synchronizing information delivering observations on activity trends promptly. Each element of the plan is thoughtfully selected to enhance the flow of data and speed up analysis.

At the start of our analysis process, for Twitter data focusing on circadian rhythms analysis, we prepared the data needed for it using the Twitter API and the Sleepwalking dataset that includes timestamps along with user location details and demographic information.

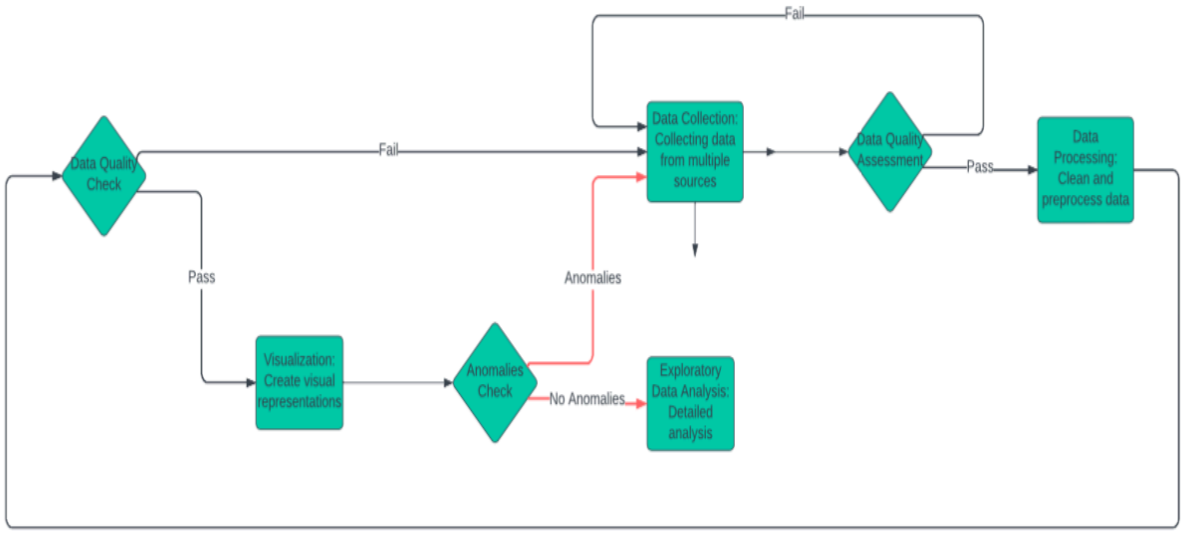

Figure A1: Flowchart illustrating the Data Science Lifecycle

We follow the data science lifecycle, to provide a solution to the problem. This is of 4 stages.

1. Data Collection:

Data can be acquired in 2 different ways. Provided datasets, which is straight forward or web scraping which is not so straight forward.

We have limited access to acquire the tweets through developer account

Web scraping techniques were implemented with the help of some libraries called Twikit and tweepy, but no tweets were obtained from it due to the newly modified regulations from X (Twitter). Below mentioned code snippets are some of the trials performed to get the tweets from twitter.

```python
import twikit
from twikit import Client, TooManyRequests
import time
from datetime import datetime
import csv
from configparser import ConfigParser
from random import randint


MINIMUM_TWEETS = 10
QUERY = 'elonmusk'


def get_tweets(tweets):
    if tweets is None:
        #* get tweets
        print(f'{datetime.now()} - Getting tweets...')
        tweets = client.search_tweet(QUERY, product='Top')
    else:
        wait_time = randint(5, 10)
        print(f'{datetime.now()} - Getting next tweets after {wait_time} seconds ...')
        time.sleep(wait_time)
        tweets = tweets.next()

    return tweets
```

Using a Python module called twikit, the code snippet is trying to retrieve tweets from Twitter. It defines a function called get_tweets that either retrieves the subsequent batch of tweets or a fresh set of tweets that match a particular query. It also has a few imports and constants that are necessary for it to function.

But there is a forbidden access error because X (Twitter) has blocked all the API's and libraries. So, we used available datasets of twitter and worked on it.

```
import tweepy

consumer_key = "S6my2HeNbWgvZoFyKpQcQMfXq" #Your API/Consumer key
consumer_secret = "t6XkyHQLAqJEeYLNC9Uwwdbtuu6Asaqh4MJdKJUkbcjR6tYMcc" #Your API/Consumer Secret Key
access_token = "1838248719726104577-fDdLCRMsqvKWCuq4sYpc2pEgnynhIH"    #Your Access token key
access_token_secret = "kHl15Pn7hPcmLs1xlQjVWi240EhBOVjlhRT9oPKGMj0tE" #Your Access token Secret key

#Pass in our twitter API authentication key
auth = tweepy.OAuth1UserHandler(
    consumer_key, consumer_secret,
    access_token, access_token_secret
)

#Instantiate the tweepy API
api = tweepy.API(auth, wait_on_rate_limit=True)


search_query = "covid19"
no_of_tweets =15


try:
    #The number of tweets we want to retrieved from the search
    tweets = api.search_tweets(q=search_query, count=no_of_tweets)

    #Pulling Some attributes from the tweet
    attributes_container = [[tweet.user.name, tweet.created_at, tweet.favorite_count, tweet.source,  tweet.text] for tweet in tweets]

    #Creation of column list to rename the columns in the dataframe
    columns = ["User", "Date Created", "Number of Likes", "Source of Tweet", "Tweet"]

    #Creation of Dataframe
    tweets_df = pd.DataFrame(attributes_container, columns=columns)
except BaseException as e:
    print('Status Failed On,',str(e))
```

```
Status Failed On, 403 Forbidden
453 - You currently have access to a subset of Twitter API v2 endpoints and limited v1.1 endpoints (e.g. media post, oauth) only. If you need access to this endpoint, you may need a different access level.
```

We have acquired jobs_sleepwalk dataset from google. This dataset has sample tweets from the year 2020.

```
#Read the file
Circadian = pd.read_csv('jobs_sleepwalk.csv')
Circadian.head() #self validation
```

	characteristic	utc_timestamp	user_hash	location
0	cco	1578017419659	c1e38841a17db91a76d618b1f52ed112	NaN
1	cco	1578034492659	aae1a10c648c1db88d888ed4041fcb32	Small Town, Florida
2	cco	1578074825657	ba78bb0e9d03438e17db85cd385de2ac	NaN
3	cco	1578108006666	c1e38841a17db91a76d618b1f52ed112	NaN
4	cco	1578193720663	aae1a10c648c1db88d888ed4041fcb32	Small Town, Florida

There are four features and 4547724 rows of data in this dataset. Characteristic feature represent the jobs and roles of the tweeters, utc_timestamp is the converted form of the timezones, User_hash is the unique ID of the users and location is the place of the users.

2. Data Processing:

The greatest focus and effort was needed to clean the data. Before any analysis can start, the data must be cleaned up, according to an exploratory run of the dataset. Variations in how devices indicate a user's position or errors in user input are examples of this dirt or noise in the data. We cleaned the data throughout the cleaning step using Python and Python libraries such as Pandas.

```
#Converting time stamp
Circadian['Timestamp'] = pd.to_datetime(Circadian['Timestamp'],unit='ms')
Circadian.head(200)
```

	Occupation	Timestamp	User ID	Location
0	cco	2020-01-03 06:54:52.659	aae1a10c648c1db88d888ed4041fcb32	Small Town, Florida
1	cco	2020-01-05 03:08:40.663	aae1a10c648c1db88d888ed4041fcb32	Small Town, Florida
2	cco	2020-01-05 06:46:11.659	87ac024dd6fdbb40cec6ca7c8e91a254	tokyo/london
3	cco	2020-01-05 10:05:26.663	82eefd56247d1c0b29078f4331e644de	Torino
4	cco	2020-01-06 10:09:36.662	82eefd56247d1c0b29078f4331e644de	Torino

This code helps in translating UTC times into a format that is readable.

```
Circadian = Circadian[~Circadian['Location'].isin(unwanted_locations)]
# 3. Remove rows containing specific unwanted symbols
unwanted_patterns = [r'\\', r'/', r'\|',r'-']  # Escape | since it's special in regex
for pattern in unwanted_patterns:
    Circadian = Circadian[~Circadian['Location'].str.contains(pattern, regex=True, na=False)]

Circadian.head(200)
Circadian.shape
```

This piece of code will eliminate all the patterns and characters which are not necessary.

Even though we tried to delete rows with improper locations, it is a time consuming and tedious task. So, to make it simpler, top places are chosen and only they are considered for the location as shown in the below snippet.

```
[37] #To get list of rows with mentioned location.
     Circadian_Filtered = Circadian[Circadian['Location'].isin(['India', 'USA', 'China', 'England', 'Ghana','Canada','Indonesia','New zealand','Ecuador',
                                                                'Columbia','Pakistan','France','Germany','Switzerland','Egypt','Italy'])]
     Circadian_Filtered = Circadian[Circadian['Location'].isin(['Andhra Pradesh','Tamilnadu','Maharashtra','Karnataka','New Delhi','Texas','California',
                                                                'Florida','Pennsylvania','New York','Belgium', 'Austria', 'Ireland','Geneva','Hawaii','Nevada','Shandong','Alaska'])]
     Circadian_Filtered = Circadian[Circadian['Location'].isin(['Delhi','Hyderabad','Bangalore','Chennai','Mumbai','Los Angeles','Chicago','Houston',
                                                                'New york City','Philadelphia','Moscow','Madrid','Barcelona', 'Beijing','Hong kong',
                                                                'Shanghai','Guangzhou','Shenzhen', 'London','France','Berlin','Zurich','Karachi',
                                                                'Lahore','Toronto','Ottawa','Vancouver','Montreal','Auckland','Wellington', 'Christchurch'])]

     Circadian_Filtered.shape
```

3. Visualization:

We used python packages like Matplotlib, Seaborn, Plotly to show the filtered data. Through this we can depict the pattern and reach some conclusions.

4. Exploratory Data Analysis (EDA):

Our aim of this project is to analyze the sleep patterns and how it affects our day-to-day life activities. In this project, we analyze this by comparing how they differ from region to region, peak hours and so on. This analysis can help us set some standards so that it does not affect mentally, physically, personally and professionally.

Key Findings

1. **Influence of Rhythms, on Twitter Usage Trends:**

The research discovered that Twitter usage showed differences in time zones which demonstrate how circadian rhythms affect user interaction on the platform These trends showcase the

influence of natural biological patterns, like waking and sleeping schedules that are shaped by regional customs, work routines and societal standards.

People living in areas with late night media habits tend to be active during hours which may disrupt their sleep patterns naturally occurring. This tendency is particularly noticeable, in settings where technology access is greater and maintaining a work life balance can be challenging.

Regional variations in activity patterns demonstrate how where you live and local customs impact when and how often people use social media platforms like regions show activity times based on work schedules and societal norms with some places having less late night engagement due to early work start times than those with more flexible schedules.

2. **Insights into social media and its effect on sleep habit:**

Late night social media use is linked to sleep disturbances as, per a study that found people who're active on Twitter when they should be sleeping might suffer from sleep deprivation leading to effects on mental health and physical and cognitive abilities.

Studies have shown that younger people and those with demanding jobs often engage in media late at night which may worsen problems related to sleep quality. This observation underscores the significance of raising awareness about the significance of developing habits to support a healthy sleep routine.

The research results offer advice for healthcare professionals and decision makers to tackle issues related to sleep health, such, as promoting campaigns promoting mindful use of social media and designing resources to assist individuals in better managing their online habits.

3. **Effect on professional efficiency:**

The disturbance of natural sleep patterns caused by using media at night doesn't just affect personal health but also has a significant impact on how productive we are, in both work and personal life situations. Studies have demonstrated that inadequate sleep resulting from night social media use can lead to decreased concentration levels and impaired decision making abilities which ultimately hinder overall effectiveness in the workplace.

Thus, the research highlights the importance of exploring how our digital behaviors impact our body's clock and the overall effects on health and efficiency.

Conclusion

The study met its goals by uncovering information about daily rhythm trends from past Twitter data analysis efforts. The team employed methods for handling and presenting data making it easy to spot behavior patterns among age groups and locations. These accomplishments highlight the promise of leveraging social media information for evaluations and set the stage for future research, in related fields.

Nonetheless the project encountered obstacles such as restricted access to data owing to limitations on Twitter's API and the intricacies linked with managing and analyzing substantial datasets effectively. These constraints highlight the necessity for strategies for obtaining data and a stronger infrastructure, for handling datasets.

The outlook presented in the study holds prospects for further growth and progress ahead of us. Automating data preparation processes could greatly enhance effectiveness and scalability while integrating data origins like social media channels or wearable health gadgets might enrich the breadth of examination. These progressions could broaden the relevance of the study across fields such as healthcare policy making and professional efficiency ultimately aiding an insight into how digital actions and human bodily cycles interact.

References

British Neuroscience Association (BNA).(n.d.). “Using social media to investigate circadian rhythms.” *British Neuroscience Association.* https://www.bna.org.uk/mediacentre/news/using-social-media-to-investigate-circadian-rhythms/ (Nov. 17, 2024).

Rizoiu, M.-A., Mishra, S., and Xie, L. (2022). "Approximating circadian rhythms using social media data." *Proceedings of the International Conference on Web and Social Media (ICWSM)*. https://ojs.aaai.org/index.php/ICWSM/article/view/22202 (Nov. 17, 2024).