CS6109 COMPILER DESIGN

B.E CSE V Q - BATCH

Analysing and predicting the behaviour of whatsapp chats using sentiment analysis

S. No.	REG. NO.	NAME
1	2019103017	ESWARAMOORTHY
2	2019103023	K JAYAKRISHNA
3	2019103038	MURALI R

ABSTRACT:

What does other person feel about us is the most frequently self-asked question, Everyone has the curiosity of what other person thinks about the other while having a conversation ,judging the other person can’t be done perfectly ,So this paper is providing a way using sentiment analysis between conversation .While chatting with other person we always have a question about our image on the other person mind. This process deals with pre-processing the data obtained from the WhatsApp chat which is exported to a server and then sentiment analysis is applied for each message and all of the messages’ sentiment is normalized from a proposed method and overall sentiment is found out.

DELIVERABLES:

INPUT: A group chat from Whatsapp.

OUTPUT: Behaviour of each message and overall sentiment of the conversation.

Objective:

The main objective of this project is to find the behaviour of the chat from whatsapp.

Problem statement:

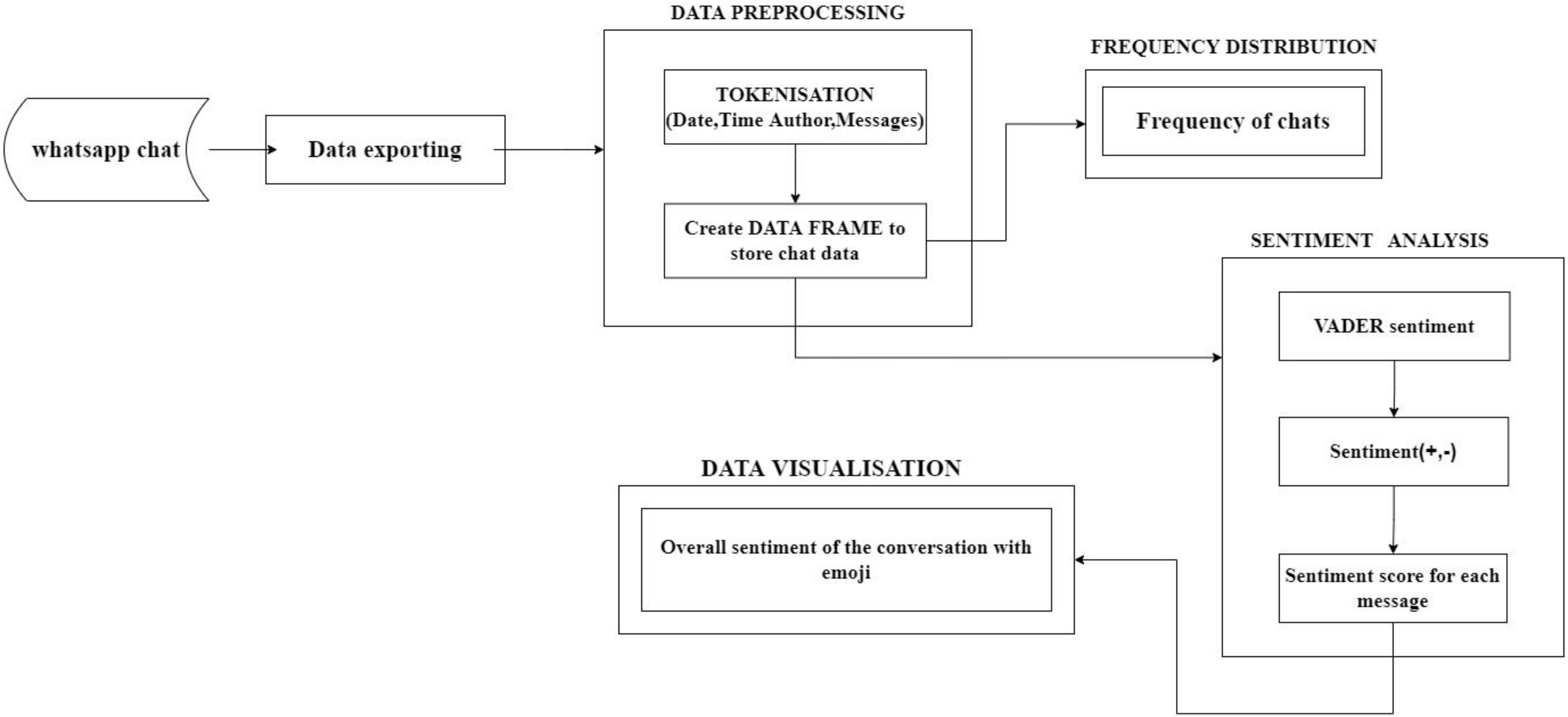
Generally,behaviour of a text can be found out easily but getting the chat from whatsapp and processing it and analysing the behaviour is quite difficult to do.

LITERATURE SURVEY:

The chat analyzer is a gateway to the unexplored field of WhatsApp Chats. It allows the analysis of bilingual users. Most work done in emotion analysis focuses on classifying the user's emotions as Positive, Negative or Neutral. The Chat Analyzer goes a step ahead and classifies these emotions into six different emotions and uses their Neutrality to weigh them. It also classifies different emojis as emojis are a popular way of expressing emotions. Our Analyzer gave 72.9% accuracy against a set of pre-classified data. A rise in hate speech, cyber-bullying, heckling and increased impudence on social media has been observed. The Chat Analyzer can be used as a tool to give a user an insight on their online behavior as they communicate with their peers. It allows users to keep a check on their emotions by analyzing them. The model described in this paper is successfully applicable to WhatsApp Chats, classifies the texts into one of 6 emotions while taking into consideration the emojis used by the person and therefore stands apart.

Contribution: we used python and its libraries nltk to data preprocessing.VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis one of the nltk libraries used here to find sentiment polarity of chats.

Block Diagram



MODULES:

1. DATA EXPORTING:

Whats app provides a feature of exporting any chat (with or without media) as a .txt file.

INPUT:whatsapp chat

OUTPUT:exported dataset of chats without media.

2. DATA PREPROCESSING:

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

Input:Dataset

Output:Author,Message ,Date Time

Create the data frame: Now, the following code will take care of using the first code chunk and placing all the data into a data frame format and changing the names of users for privacy.

Frequency distribution:

We can now analyze the information we have, some general stats we can look into are the total amount of messages that have been shared with the group, how many media items and how many links in total.

Output:Frequency of chats

3.SENTIMENT ANALYSIS:

The process of ‘computationally’ determining whether a piece of writing is positive, negative or neutral.VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative.

Input: Message Separated from data Processing.

Output: Polarity score for each chat.

4.DATA VISUALISATION:

At last,the data analysed is displayed with all the details we got from the whatsapp.

Output:overall behaviour of the conversation.

DATASET DESCRIPTION:

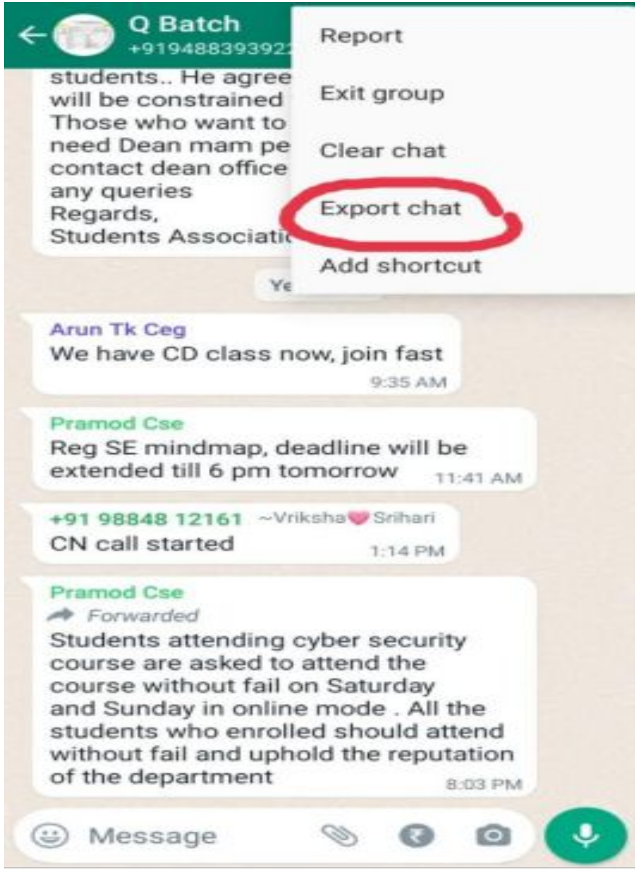
WhatsApp provides us the feature of exporting chats, so let's export the chat and save the file.Text file containing exported chat from whatsapp is used as dataset.We omit the media files during exporting.

Chat.txt

```
11/20/21, 12:30 PM - +91 98848 12161: Hostellers: if you have major issues, ma'am said
you can go to KP and write the test in person
11/20/21, 12:45 PM - Pramod Cse: Due to network issues no class today.
11/20/21, 12:45 PM - Pramod Cse: From se ma'am
11/20/21, 4:29 PM - Arun Tk Ceg: <Media omitted>
11/20/21, 4:29 PM - Arun Tk Ceg: 20 jan we have end sem exam
11/20/21, 4:29 PM - Arun Tk Ceg: Dec 21 is the last working day
11/20/21, 4:29 PM - Arun Tk Ceg: After that no class
11/20/21, 4:31 PM - Arun Tk Ceg: Classes in online mode for us, no change in that so far
11/20/21, 6:50 PM - Vishnu Priya: Grades for sem 4 updated on acoe
11/22/21, 1:42 PM - Vishnu Priya: Closes for CD lab has been removed
11/22/21, 1:42 PM - Vishnu Priya: Spot*
11/22/21, 2:21 PM - Arun Tk Ceg: Dear student, those who are not submit the scholarship
application of BC/MBC-(regular course only)/SC/ST (R & SS)for 2021-2022 submit on or
before 30.11.2021.
```

RESULT IMPLEMENTATION:

1)DATA EXPORTING



students.. He agree
will be constrained
Those who want to
need Dean mam pe
contact dean office
any queries
Regards,
Students Associati

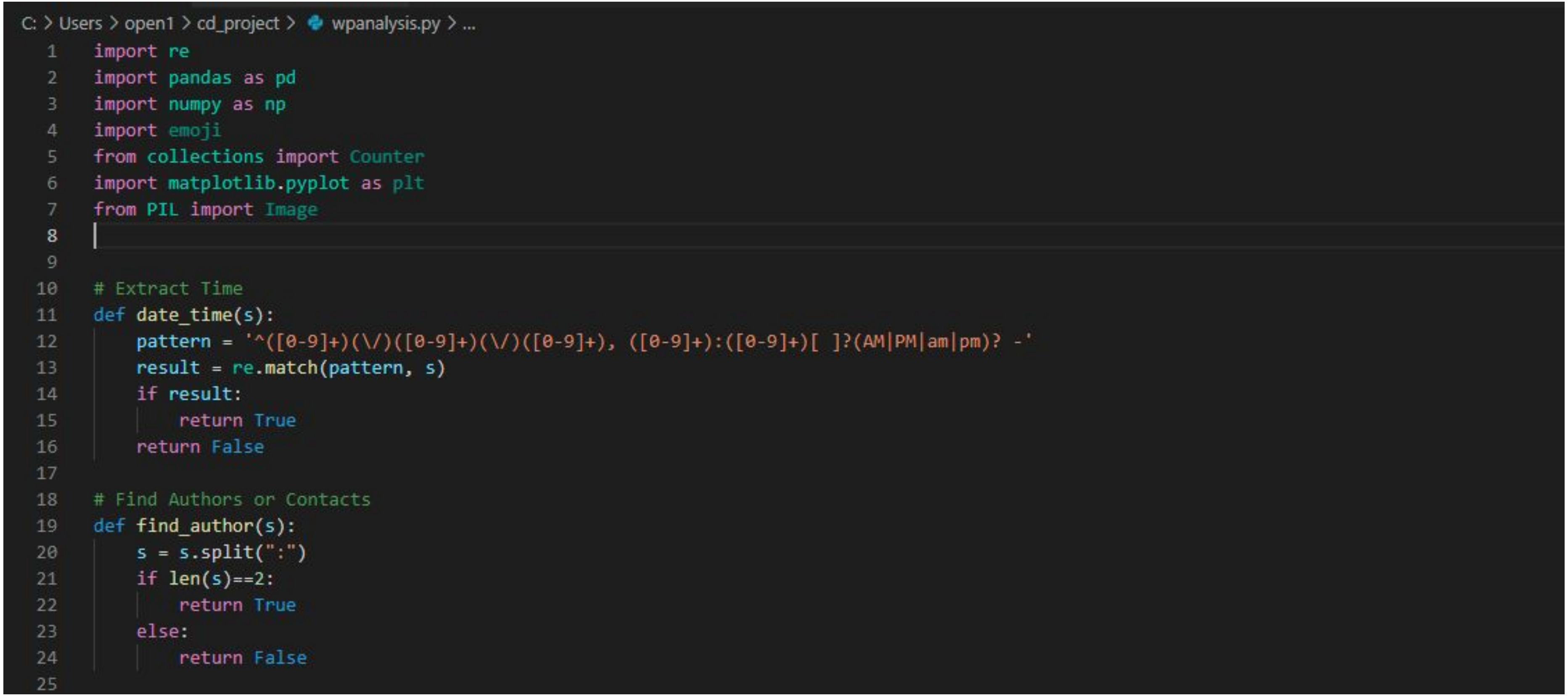
Arun Tk Ceg
We have CD class now, join fast 9:35 AM

Pramod Cse
Reg SE mindmap, deadline will be extended till 6 pm tomorrow 11:41 AM

+91 98848 12161 ~Vriksha💗Srihari
CN call started 1:14 PM

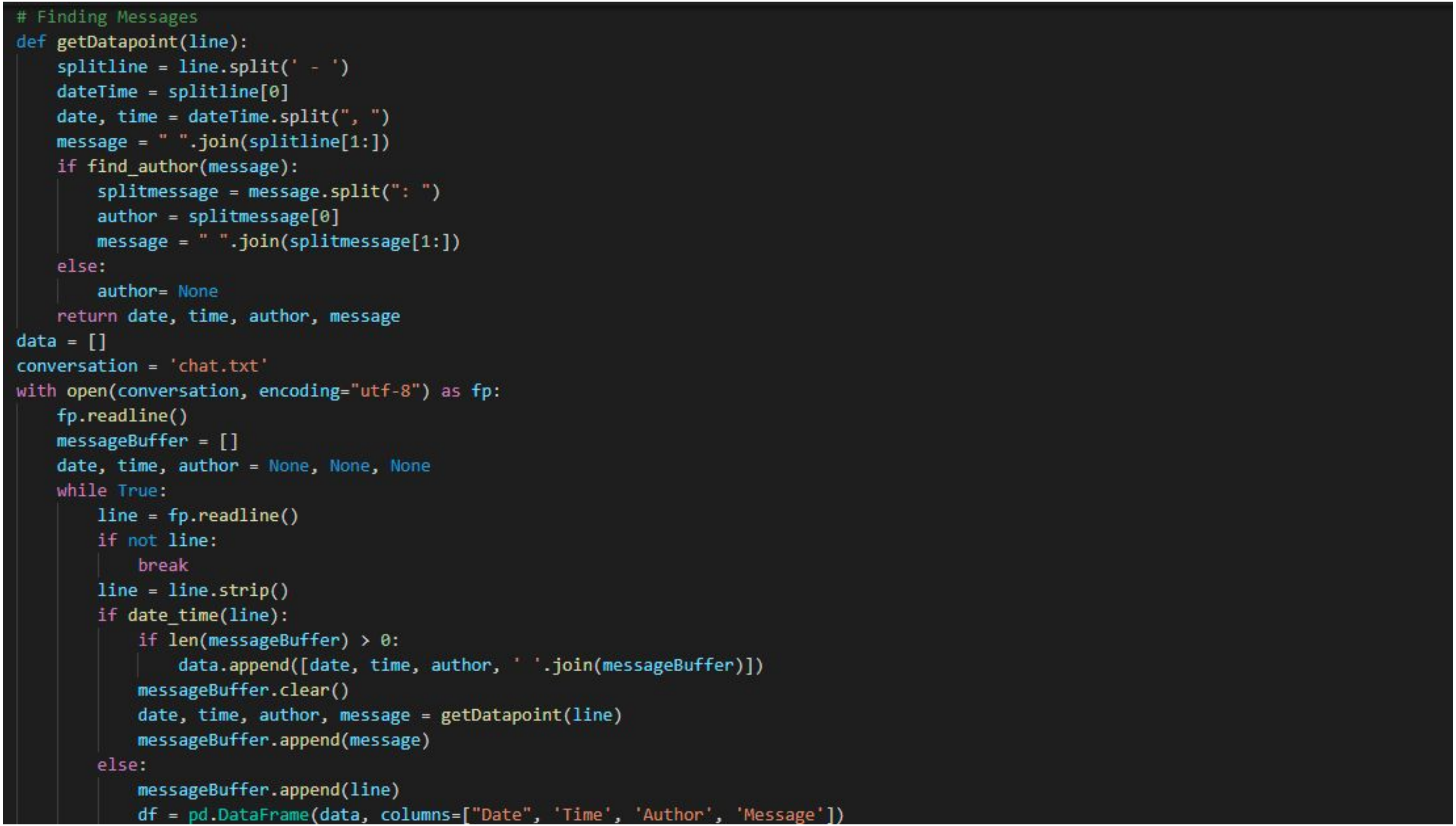
Pramod Cse
Forwarded
Students attending cyber security course are asked to attend the course without fail on Saturday and Sunday in online mode . All the students who enrolled should attend without fail and uphold the reputation of the department 8:03 PM

DATA PREPROCESSING

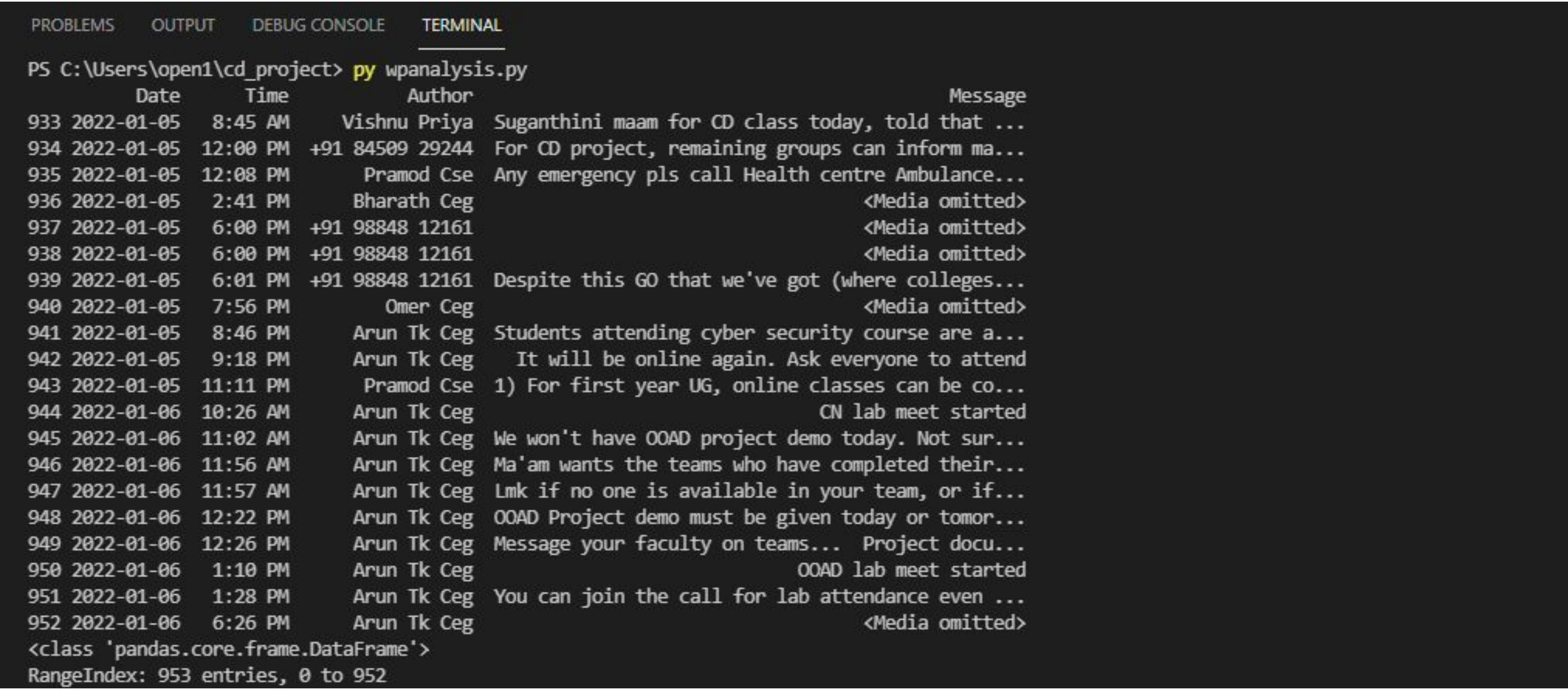


C: > Users > open1 > cd_project > wpanalysis.py > ...

```python
import re
import pandas as pd
import numpy as np
import emoji
from collections import Counter
import matplotlib.pyplot as plt
from PIL import Image


# Extract Time
def date_time(s):
    pattern = '^([0-9]+)(\/)([0-9]+)(\/)([0-9]+), ([0-9]+):([0-9]+)[ ]?(AM|PM|am|pm)? -'
    result = re.match(pattern, s)
    if result:
        return True
    return False

# Find Authors or Contacts
def find_author(s):
    s = s.split(":")
    if len(s)==2:
        return True
    else:
        return False

```

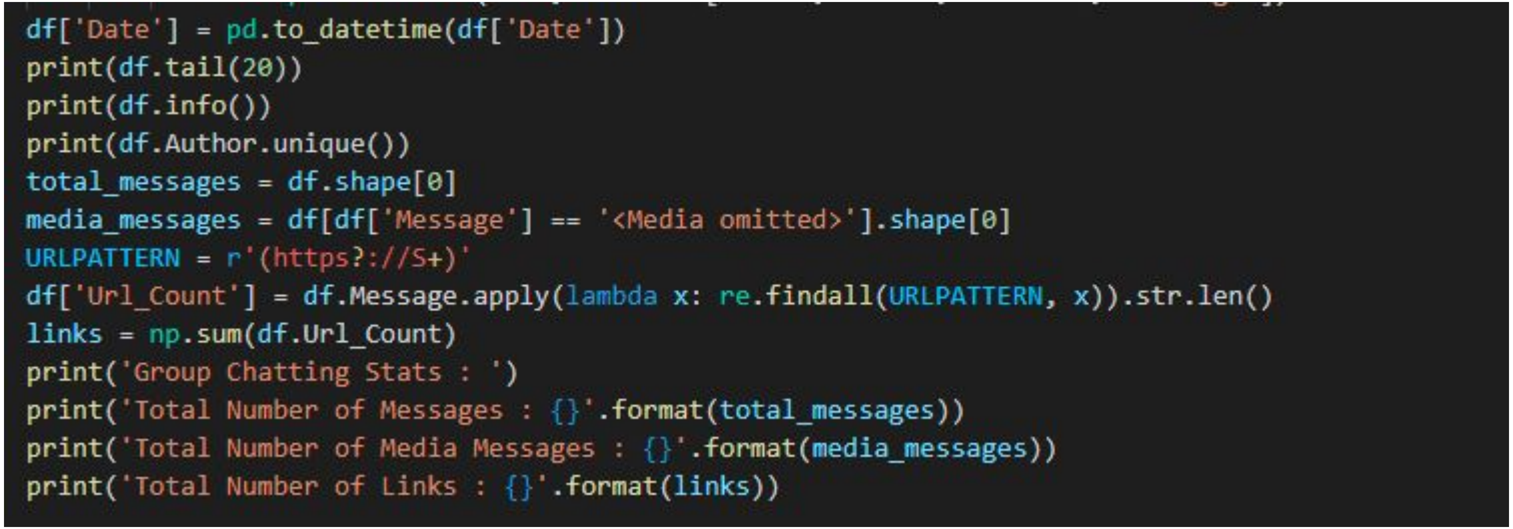


```
# Finding Messages
def getDatapoint(line):
    splitline = line.split(' - ')
    dateTime = splitline[0]
    date, time = dateTime.split(", ")
    message = " ".join(splitline[1:])
    if find_author(message):
        splitmessage = message.split(": ")
        author = splitmessage[0]
        message = " ".join(splitmessage[1:])
    else:
        author= None
    return date, time, author, message
data = []
conversation = 'chat.txt'
with open(conversation, encoding="utf-8") as fp:
    fp.readline()
    messageBuffer = []
    date, time, author = None, None, None
    while True:
        line = fp.readline()
        if not line:
            break
        line = line.strip()
        if date_time(line):
            if len(messageBuffer) > 0:
                data.append([date, time, author, ' '.join(messageBuffer)])
            messageBuffer.clear()
            date, time, author, message = getDatapoint(line)
            messageBuffer.append(message)
        else:
            messageBuffer.append(line)
            df = pd.DataFrame(data, columns=["Date", 'Time', 'Author', 'Message'])
```

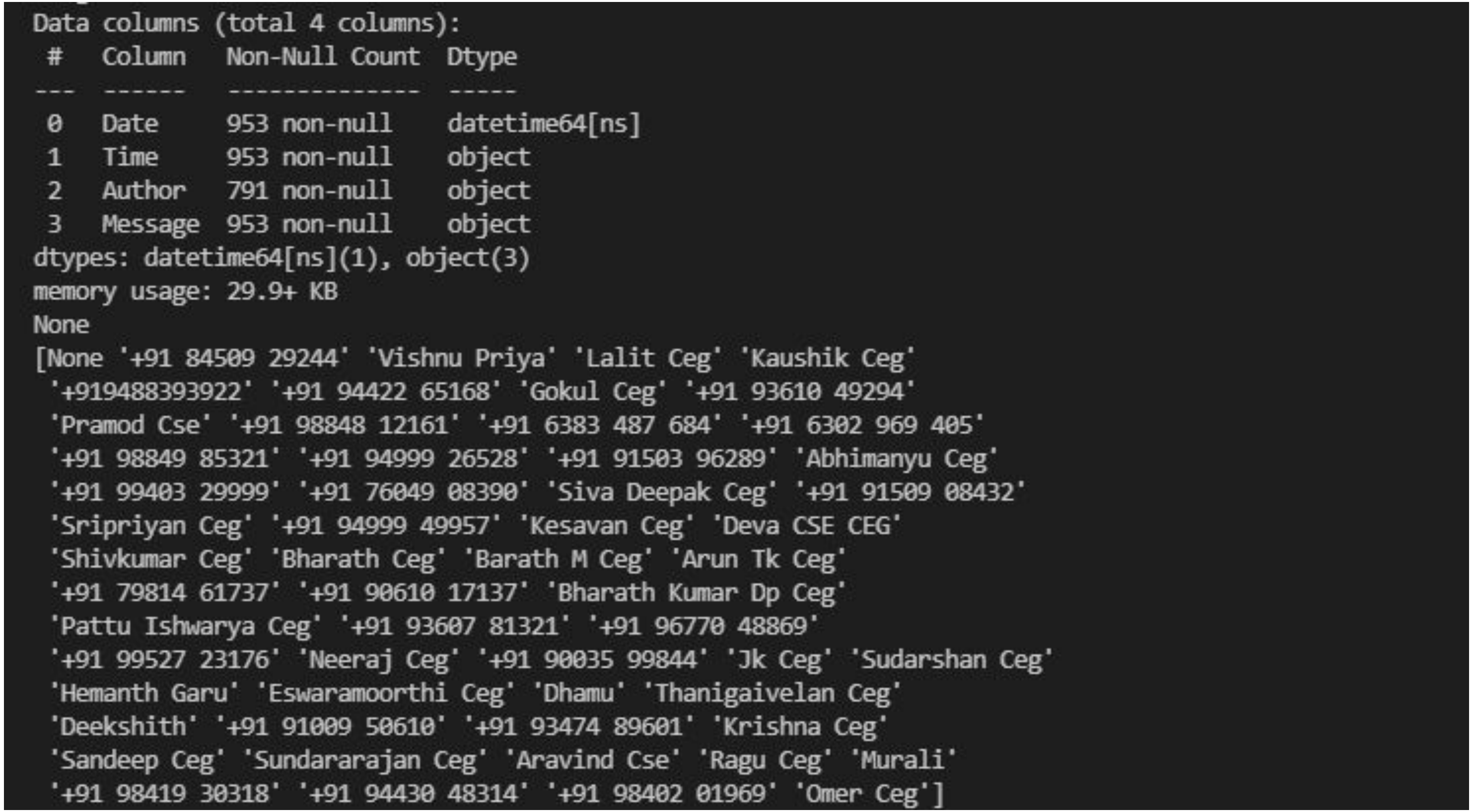


```
PS C:\Users\open1\cd_project> py wpanalysis.py
          Date      Time            Author                                            Message
933 2022-01-05   8:45 AM      Vishnu Priya  Suganthini maam for CD class today, told that ...
934 2022-01-05  12:00 PM  +91 84509 29244  For CD project, remaining groups can inform ma...
935 2022-01-05  12:08 PM       Pramod Cse  Any emergency pls call Health centre Ambulance...
936 2022-01-05   2:41 PM      Bharath Ceg                                     <Media omitted>
937 2022-01-05   6:00 PM  +91 98848 12161                                     <Media omitted>
938 2022-01-05   6:00 PM  +91 98848 12161                                     <Media omitted>
939 2022-01-05   6:01 PM  +91 98848 12161  Despite this GO that we've got (where colleges...
940 2022-01-05   7:56 PM         Omer Ceg                                     <Media omitted>
941 2022-01-05   8:46 PM      Arun Tk Ceg  Students attending cyber security course are a...
942 2022-01-05   9:18 PM      Arun Tk Ceg    It will be online again. Ask everyone to attend
943 2022-01-05  11:11 PM       Pramod Cse  1) For first year UG, online classes can be co...
944 2022-01-06  10:26 AM      Arun Tk Ceg                                  CN lab meet started
945 2022-01-06  11:02 AM      Arun Tk Ceg  We won't have OOAD project demo today. Not sur...
946 2022-01-06  11:56 AM      Arun Tk Ceg  Ma'am wants the teams who have completed their...
947 2022-01-06  11:57 AM      Arun Tk Ceg  Lmk if no one is available in your team, or if...
948 2022-01-06  12:22 PM      Arun Tk Ceg  OOAD Project demo must be given today or tomor...
949 2022-01-06  12:26 PM      Arun Tk Ceg  Message your faculty on teams...  Project docu...
950 2022-01-06   1:10 PM      Arun Tk Ceg                                OOAD lab meet started
951 2022-01-06   1:28 PM      Arun Tk Ceg  You can join the call for lab attendance even ...
952 2022-01-06   6:26 PM      Arun Tk Ceg                                     <Media omitted>
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 953 entries, 0 to 952
```

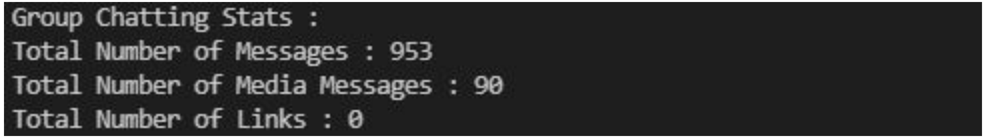
FREQUENCY DISTRIBUTION



```
df['Date'] = pd.to_datetime(df['Date'])
print(df.tail(20))
print(df.info())
print(df.Author.unique())
total_messages = df.shape[0]
media_messages = df[df['Message'] == '<Media omitted>'].shape[0]
URLPATTERN = r'(https?://S+)'
df['Url_Count'] = df.Message.apply(lambda x: re.findall(URLPATTERN, x)).str.len()
links = np.sum(df.Url_Count)
print('Group Chatting Stats : ')
print('Total Number of Messages : {}'.format(total_messages))
print('Total Number of Media Messages : {}'.format(media_messages))
print('Total Number of Links : {}'.format(links))
```

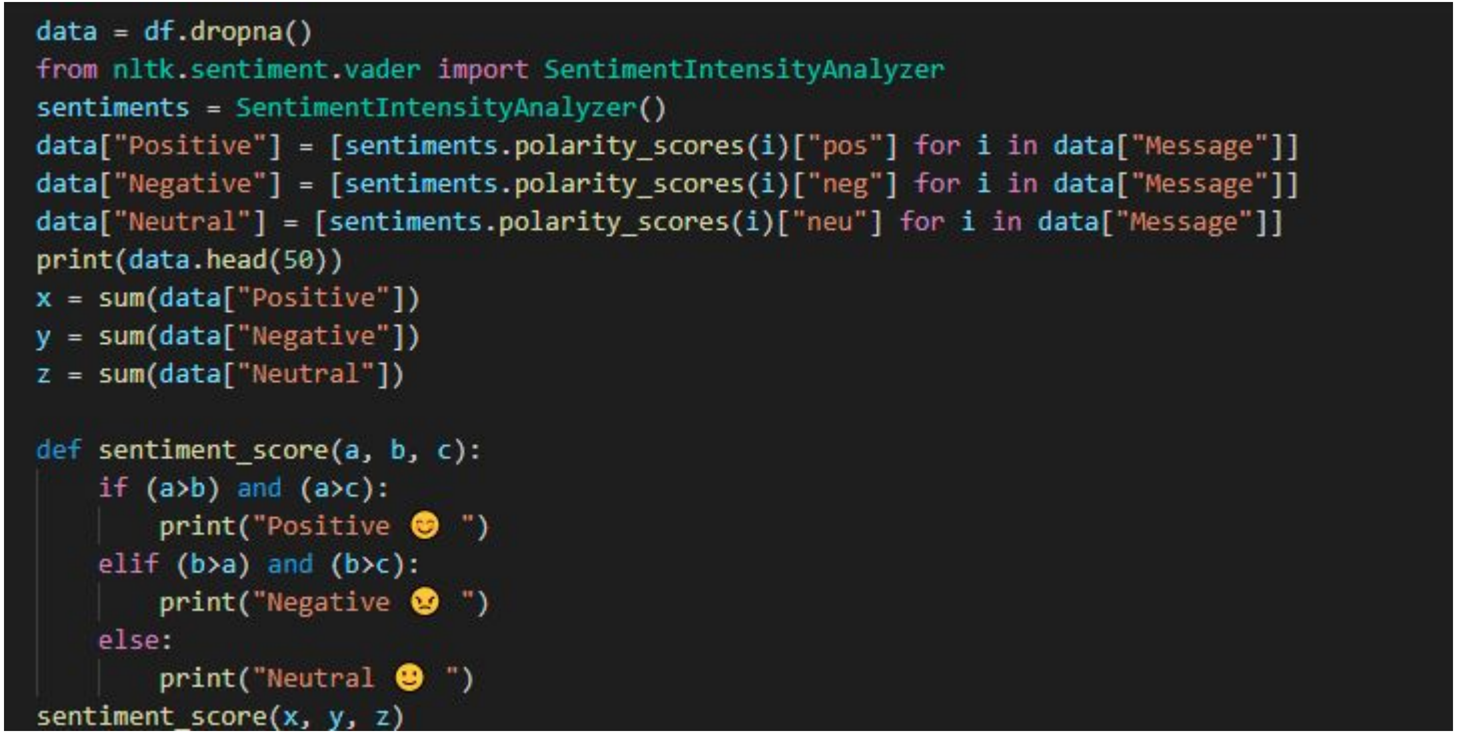


```
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   Date     953 non-null    datetime64[ns]
 1   Time     953 non-null    object
 2   Author   791 non-null    object
 3   Message  953 non-null    object
dtypes: datetime64[ns](1), object(3)
memory usage: 29.9+ KB
None
[None '+91 84509 29244' 'Vishnu Priya' 'Lalit Ceg' 'Kaushik Ceg'
 '+919488393922' '+91 94422 65168' 'Gokul Ceg' '+91 93610 49294'
 'Pramod Cse' '+91 98848 12161' '+91 6383 487 684' '+91 6302 969 405'
 '+91 98849 85321' '+91 94999 26528' '+91 91503 96289' 'Abhimanyu Ceg'
 '+91 99403 29999' '+91 76049 08390' 'Siva Deepak Ceg' '+91 91509 08432'
 'Sripriyan Ceg' '+91 94999 49957' 'Kesavan Ceg' 'Deva CSE CEG'
 'Shivkumar Ceg' 'Bharath Ceg' 'Barath M Ceg' 'Arun Tk Ceg'
 '+91 79814 61737' '+91 90610 17137' 'Bharath Kumar Dp Ceg'
 'Pattu Ishwarya Ceg' '+91 93607 81321' '+91 96770 48869'
 '+91 99527 23176' 'Neeraj Ceg' '+91 90035 99844' 'Jk Ceg' 'Sudarshan Ceg'
 'Hemanth Garu' 'Eswaramoorthi Ceg' 'Dhamu' 'Thanigaivelan Ceg'
 'Deekshith' '+91 91009 50610' '+91 93474 89601' 'Krishna Ceg'
 'Sandeep Ceg' 'Sundararajan Ceg' 'Aravind Cse' 'Ragu Ceg' 'Murali'
 '+91 98419 30318' '+91 94430 48314' '+91 98402 01969' 'Omer Ceg']
```

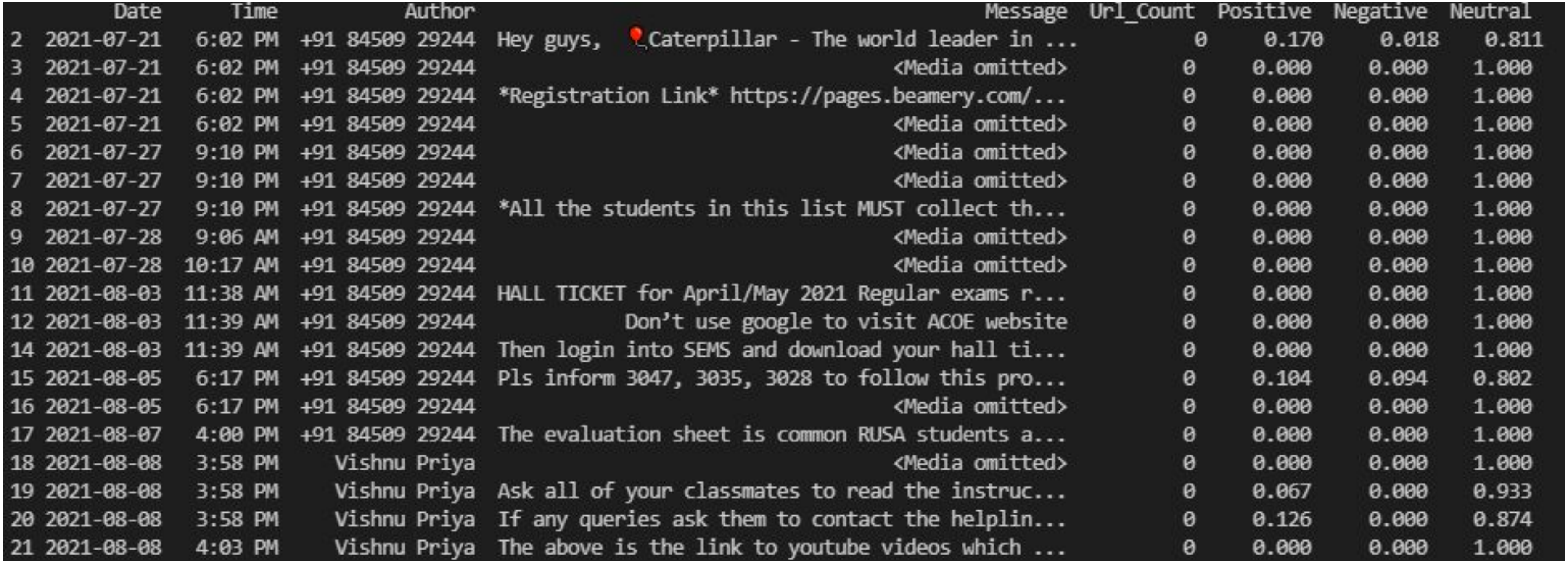


```
Group Chatting Stats :
Total Number of Messages : 953
Total Number of Media Messages : 90
Total Number of Links : 0
```

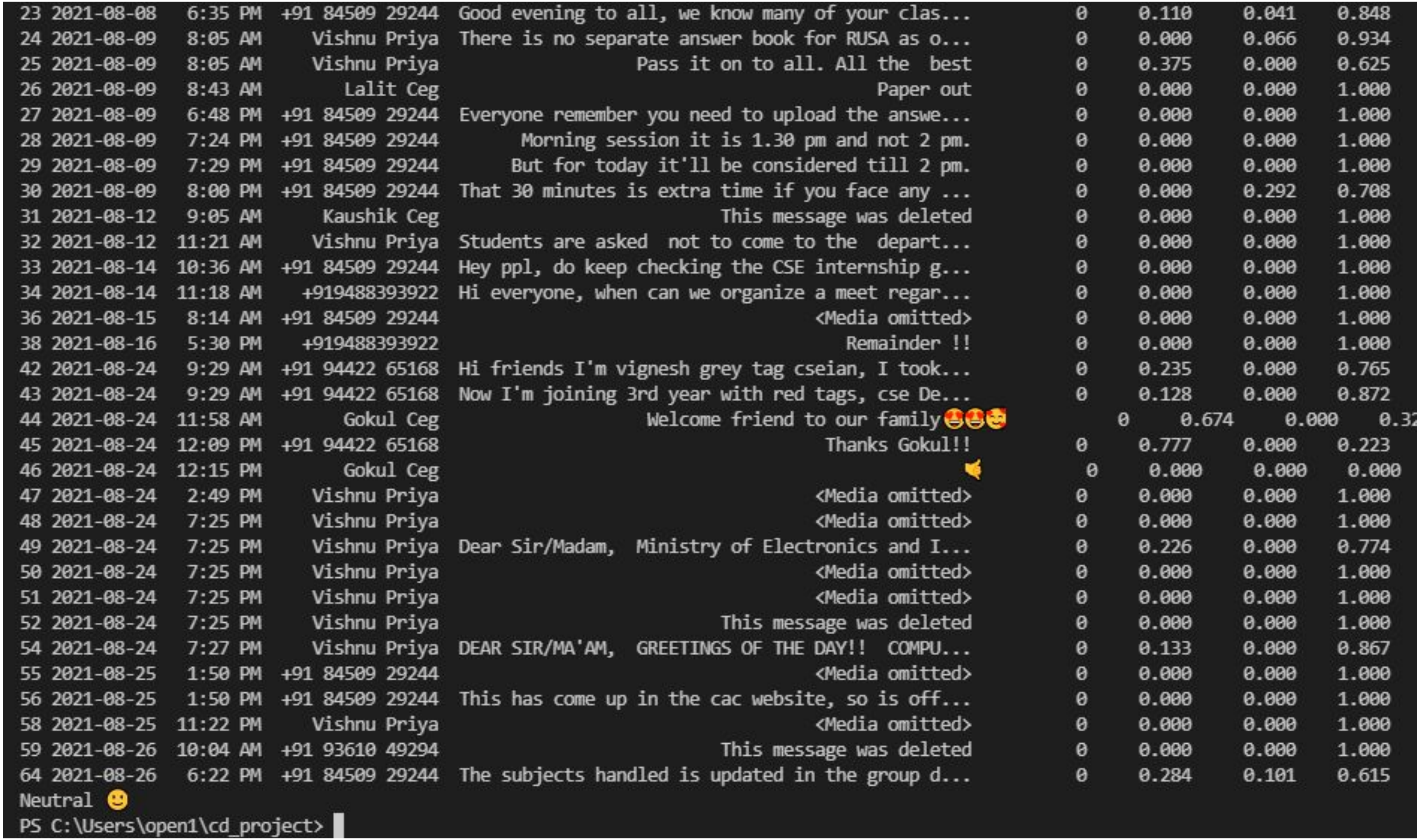
SENTIMENT ANALYSIS and DATA VISUALISATION



```python
data = df.dropna()
from nltk.sentiment.vader import SentimentIntensityAnalyzer
sentiments = SentimentIntensityAnalyzer()
data["Positive"] = [sentiments.polarity_scores(i)["pos"] for i in data["Message"]]
data["Negative"] = [sentiments.polarity_scores(i)["neg"] for i in data["Message"]]
data["Neutral"] = [sentiments.polarity_scores(i)["neu"] for i in data["Message"]]
print(data.head(50))
x = sum(data["Positive"])
y = sum(data["Negative"])
z = sum(data["Neutral"])

def sentiment_score(a, b, c):
    if (a>b) and (a>c):
        print("Positive 😊 ")
    elif (b>a) and (b>c):
        print("Negative 😠 ")
    else:
        print("Neutral 🙂 ")
sentiment_score(x, y, z)
```



```
         Date      Time          Author                                    Message  Url_Count  Positive  Negative  Neutral
2  2021-07-21   6:02 PM  +91 84509 29244  Hey guys,  📍Caterpillar - The world leader in ...          0     0.170     0.018    0.811
3  2021-07-21   6:02 PM  +91 84509 29244                                  <Media omitted>          0     0.000     0.000    1.000
4  2021-07-21   6:02 PM  +91 84509 29244  *Registration Link* https://pages.beamery.com/...          0     0.000     0.000    1.000
5  2021-07-21   6:02 PM  +91 84509 29244                                  <Media omitted>          0     0.000     0.000    1.000
6  2021-07-27   9:10 PM  +91 84509 29244                                  <Media omitted>          0     0.000     0.000    1.000
7  2021-07-27   9:10 PM  +91 84509 29244                                  <Media omitted>          0     0.000     0.000    1.000
8  2021-07-27   9:10 PM  +91 84509 29244  *All the students in this list MUST collect th...          0     0.000     0.000    1.000
9  2021-07-28   9:06 AM  +91 84509 29244                                  <Media omitted>          0     0.000     0.000    1.000
10 2021-07-28  10:17 AM  +91 84509 29244                                  <Media omitted>          0     0.000     0.000    1.000
11 2021-08-03  11:38 AM  +91 84509 29244  HALL TICKET for April/May 2021 Regular exams r...          0     0.000     0.000    1.000
12 2021-08-03  11:39 AM  +91 84509 29244                Don’t use google to visit ACOE website          0     0.000     0.000    1.000
14 2021-08-03  11:39 AM  +91 84509 29244  Then login into SEMS and download your hall ti...          0     0.000     0.000    1.000
15 2021-08-05   6:17 PM  +91 84509 29244  Pls inform 3047, 3035, 3028 to follow this pro...          0     0.104     0.094    0.802
16 2021-08-05   6:17 PM  +91 84509 29244                                  <Media omitted>          0     0.000     0.000    1.000
17 2021-08-07   4:00 PM  +91 84509 29244  The evaluation sheet is common RUSA students a...          0     0.000     0.000    1.000
18 2021-08-08   3:58 PM     Vishnu Priya                                  <Media omitted>          0     0.000     0.000    1.000
19 2021-08-08   3:58 PM     Vishnu Priya  Ask all of your classmates to read the instruc...          0     0.067     0.000    0.933
20 2021-08-08   3:58 PM     Vishnu Priya  If any queries ask them to contact the helplin...          0     0.126     0.000    0.874
21 2021-08-08   4:03 PM     Vishnu Priya  The above is the link to youtube videos which ...          0     0.000     0.000    1.000
```



```
23 2021-08-08   6:35 PM +91 84509 29244  Good evening to all, we know many of your clas...      0    0.110    0.041    0.848
24 2021-08-09   8:05 AM     Vishnu Priya  There is no separate answer book for RUSA as o...      0    0.000    0.066    0.934
25 2021-08-09   8:05 AM     Vishnu Priya                      Pass it on to all. All the  best      0    0.375    0.000    0.625
26 2021-08-09   8:43 AM        Lalit Ceg                                             Paper out      0    0.000    0.000    1.000
27 2021-08-09   6:48 PM +91 84509 29244  Everyone remember you need to upload the answe...      0    0.000    0.000    1.000
28 2021-08-09   7:24 PM +91 84509 29244        Morning session it is 1.30 pm and not 2 pm.      0    0.000    0.000    1.000
29 2021-08-09   7:29 PM +91 84509 29244       But for today it'll be considered till 2 pm.      0    0.000    0.000    1.000
30 2021-08-09   8:00 PM +91 84509 29244  That 30 minutes is extra time if you face any ...      0    0.000    0.292    0.708
31 2021-08-12   9:05 AM     Kaushik Ceg                            This message was deleted      0    0.000    0.000    1.000
32 2021-08-12  11:21 AM     Vishnu Priya  Students are asked  not to come to the  depart...      0    0.000    0.000    1.000
33 2021-08-14  10:36 AM +91 84509 29244  Hey ppl, do keep checking the CSE internship g...      0    0.000    0.000    1.000
34 2021-08-14  11:18 AM    +919488393922  Hi everyone, when can we organize a meet regar...      0    0.000    0.000    1.000
36 2021-08-15   8:14 AM +91 84509 29244                                     <Media omitted>      0    0.000    0.000    1.000
38 2021-08-16   5:30 PM    +919488393922                                        Remainder !!      0    0.000    0.000    1.000
42 2021-08-24   9:29 AM +91 94422 65168  Hi friends I'm vignesh grey tag cseian, I took...      0    0.235    0.000    0.765
43 2021-08-24   9:29 AM +91 94422 65168  Now I'm joining 3rd year with red tags, cse De...      0    0.128    0.000    0.872
44 2021-08-24  11:58 AM        Gokul Ceg                  Welcome friend to our family😍😍🥰      0    0.674    0.000    0.32
45 2021-08-24  12:09 PM +91 94422 65168                                      Thanks Gokul!!      0    0.777    0.000    0.223
46 2021-08-24  12:15 PM        Gokul Ceg                                                   🤙      0    0.000    0.000    0.000
47 2021-08-24   2:49 PM     Vishnu Priya                                     <Media omitted>      0    0.000    0.000    1.000
48 2021-08-24   7:25 PM     Vishnu Priya                                     <Media omitted>      0    0.000    0.000    1.000
49 2021-08-24   7:25 PM     Vishnu Priya  Dear Sir/Madam,  Ministry of Electronics and I...      0    0.226    0.000    0.774
50 2021-08-24   7:25 PM     Vishnu Priya                                     <Media omitted>      0    0.000    0.000    1.000
51 2021-08-24   7:25 PM     Vishnu Priya                                     <Media omitted>      0    0.000    0.000    1.000
52 2021-08-24   7:25 PM     Vishnu Priya                            This message was deleted      0    0.000    0.000    1.000
54 2021-08-24   7:27 PM     Vishnu Priya  DEAR SIR/MA'AM,  GREETINGS OF THE DAY!!  COMPU...      0    0.133    0.000    0.867
55 2021-08-25   1:50 PM +91 84509 29244                                     <Media omitted>      0    0.000    0.000    1.000
56 2021-08-25   1:50 PM +91 84509 29244  This has come up in the cac website, so is off...      0    0.000    0.000    1.000
58 2021-08-25  11:22 PM     Vishnu Priya                                     <Media omitted>      0    0.000    0.000    1.000
59 2021-08-26  10:04 AM +91 93610 49294                            This message was deleted      0    0.000    0.000    1.000
64 2021-08-26   6:22 PM +91 84509 29244  The subjects handled is updated in the group d...      0    0.284    0.101    0.615
Neutral 🙂
PS C:\Users\open1\cd_project>
```

Conclusion:

In conclusion we can argue for days stating this is a breach of privacy to analyse a person chat and process the emotions but when we see the brighter side of the project ,we can help people develop their character by helping them suggesting proper vocabulary to guide them in a proper path to reach a sustainable position with the other end person in terms of relations, business agreements and many more .Also this requires huge amount of data to master the art emotion is relative to each person and varies so the system has to be trained in a particular way to keep track of the context and past history while determining the sentiment

REFERENCES:

1)Kaur, R. (2019). Insight to Emotional tones in WhatsApp Through Sentiment Analysis. IJRAR.

2)Winarko, E., &Cherid, A. (2017, November). Recognizing the sarcastic statement on WhatsApp Group with Indonesian language text.In 2017 International Conference on Broadband Communication, Wireless Sensors and Powering (BCWSP) (pp. 1-6).IEEE.

3)Waterloo, S. F., Baumgartner, S. E., Peter, J., & Valkenburg, P. M. (2018). Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp. new media & society, 20(5), 1813-1831.

4)Preotiuc-Pietro, D., et al.: Trendminer: an architecture for real time analysis of social media text. In: Proceedings of the Workshop on Real-Time Analysis and Mining of Social Streams (2012)

6)Gupta, P., & Nene, M. J. (2017, March). Analysis of Text Messages in Social Media to Investigate CyberPsycho Attack.In International Conference on Information and Communication Technology for Intelligent Systems (pp. 581-587).Springer, Cham.