Sentiment analysis

1. **The name of student**:

   Yerkebulan Bauyrzhanov

Objective

The goal of this programming exercise is to demonstrate your ability to design a solution to a problem and implement this solution in Python using software engineering best practices. The specific task will be to collect a dataset and perform first analysis on it. To build this application, you will crawl Telegram messages, filter non-English messages, and compute the average sentiment over time.

Pre-process the data. Remove non-English messages. From these, keep only messages that mention either "SHIB" or "DOGE." Use the tqdm package to display progress on the terminal. Use PEP8 Style Guide for your python code.

In [1]:
```
#import necessary libraries
import json
import pandas as pd
import re
from tqdm import tqdm
from textblob import TextBlob
import plotly.express as px
from langdetect import detect
```

In [2]:
```
# Opening JSON file
f = open('result.json', encoding="utf8")
# returns JSON object as
# a dictionary
data = json.load(f)
messages = data['messages']
# Iterating through the json
# list
listo = []
for i in tqdm(messages):
    s = i['text']
    if 'SHIB' in s or 'DOGE.' in s:
        listo.append(i)
# Closing file
#print(listo)
f.close()
```

```
100%|██████████████████████████████████████████████████████████████████████████| 47231/47231
[00:00<00:00, 808658.73it/s]
```

In [3]:
```
# Remove Non-English messages
# using search() to get only those strings with alphabets
clean_data = []
for i in tqdm(listo):
    s = i['text']
    if detect(s) == 'en':
        clean_data.append(i)
df = pd.DataFrame.from_dict(clean_data)
```

```
100%|██████████████████████████████████████████████████████████████████████████| 31
4/314 [00:05<00:00, 59.10it/s]
```

In [4]:
```
# leave only necessary columns
df = df[['date', 'text']]
```

In [5]:
```
df
```

Out[5]:

	date	text
0	2021-05-05T00:07:46	OMG, how can people suggest DOGE... Come on, h...
1	2021-05-08T08:20:08	Holy Cow CDC listed SHIB! 😚♡
2	2021-05-08T08:29:06	Why can't buy SHIB
3	2021-05-08T08:32:55	Is the SHIBA INU available to buy in the USA?
4	2021-05-08T08:34:32	Is the SHIBA INU available to buy in the USA?
...	...	...
270	2021-05-14T00:57:36	Need to put in CRO and in 45 days you get some...
271	2021-05-14T08:14:20	Hello, I have a question. What is the differen...
272	2021-05-14T10:45:14	hope I can get some help here. attempting to b...
273	2021-05-14T11:16:09	if anyone has the issue I was having, not bein...
274	2021-05-14T23:27:04	new coin SHIB is good or not

275 rows × 2 columns

In [6]:
```
#create a functiom to clean data from emoji
def remove_emoji(string):
    emoji_pattern = re.compile("["
                               u"\U0001F600-\U0001F64F"  # emoticons
                               u"\U0001F300-\U0001F5FF"  # symbols & pictographs
                               u"\U0001F680-\U0001F6FF"  # transport & map symbols
                               u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
                               u"\U00002500-\U00002BEF"  # chinese char
                               u"\U00002702-\U000027B0"
                               u"\U00002702-\U000027B0"
                               u"\U000024C2-\U0001F251"
                               u"\U0001f926-\U0001f937"
                               u"\U00010000-\U0010ffff"
                               u"\u2640-\u2642"
                               u"\u2600-\u2B55"
                               u"\u200d"
                               u"\u23cf"
                               u"\u23e9"
                               u"\u231a"
                               u"\ufe0f"  # dingbats
                               u"\u3030"
                               "]+", flags=re.UNICODE)
    return emoji_pattern.sub(r'', string)
```

In [7]:
```
# remove emoji from data
df['text'] = df['text'].apply(remove_emoji)
```

In [8]:
```
df
```

Out[8]:

	date	text
0	2021-05-05T00:07:46	OMG, how can people suggest DOGE... Come on, h...
1	2021-05-08T08:20:08	Holy Cow CDC listed SHIB!
2	2021-05-08T08:29:06	Why can't buy SHIB
3	2021-05-08T08:32:55	Is the SHIBA INU available to buy in the USA?
4	2021-05-08T08:34:32	Is the SHIBA INU available to buy in the USA?
...	...	...
270	2021-05-14T00:57:36	Need to put in CRO and in 45 days you get some...
271	2021-05-14T08:14:20	Hello, I have a question. What is the differen...
272	2021-05-14T10:45:14	hope I can get some help here. attempting to b...
273	2021-05-14T11:16:09	if anyone has the issue I was having, not bein...
274	2021-05-14T23:27:04	new coin SHIB is good or not

275 rows × 2 columns

Compute the sentiment of each message.

In [9]:
```
# create a function to get the polarity
def getPolarity(text):
    return TextBlob(text).sentiment.polarity
```

In [10]:
```
# create a new column
df['Polarity'] = df['text'].apply(getPolarity)
```

In [11]:
```
df
```

Out[11]:

	date	text	Polarity
0	2021-05-05T00:07:46	OMG, how can people suggest DOGE... Come on, h...	-0.125000
1	2021-05-08T08:20:08	Holy Cow CDC listed SHIB!	-0.166667
2	2021-05-08T08:29:06	Why can't buy SHIB	0.000000
3	2021-05-08T08:32:55	Is the SHIBA INU available to buy in the USA?	0.400000
4	2021-05-08T08:34:32	Is the SHIBA INU available to buy in the USA?	0.400000
...	...	...	...
270	2021-05-14T00:57:36	Need to put in CRO and in 45 days you get some...	0.000000
271	2021-05-14T08:14:20	Hello, I have a question. What is the differen...	0.000000
272	2021-05-14T10:45:14	hope I can get some help here. attempting to b...	0.000000
273	2021-05-14T11:16:09	if anyone has the issue I was having, not bein...	0.500000
274	2021-05-14T23:27:04	new coin SHIB is good or not	0.418182

275 rows × 3 columns

In [12]:
```
# Changing object type column to datetime
```

```
df['date'] = pd.to_datetime(df.date)

# Creating new column with just the date
df['date'] = df['date'].dt.date
```

In [13]:
```
df
```

Out[13]:

	date	text	Polarity
0	2021-05-05	OMG, how can people suggest DOGE... Come on, h...	-0.125000
1	2021-05-08	Holy Cow CDC listed SHIB!	-0.166667
2	2021-05-08	Why can't buy SHIB	0.000000
3	2021-05-08	Is the SHIBA INU available to buy in the USA?	0.400000
4	2021-05-08	Is the SHIBA INU available to buy in the USA?	0.400000
...	...	...	...
270	2021-05-14	Need to put in CRO and in 45 days you get some...	0.000000
271	2021-05-14	Hello, I have a question. What is the differen...	0.000000
272	2021-05-14	hope I can get some help here. attempting to b...	0.000000
273	2021-05-14	if anyone has the issue I was having, not bein...	0.500000
274	2021-05-14	new coin SHIB is good or not	0.418182

275 rows × 3 columns

Plot the number of messages per day and the average sentiment per day using the plotly visualization library.

In [14]:
```
# Get the average sentiment per day
df2 = df.groupby(df['date']).mean()
```

In [15]:
```
# reset index
df2 = df2.reset_index()
```

In [16]:
```
df2
```

Out[16]:

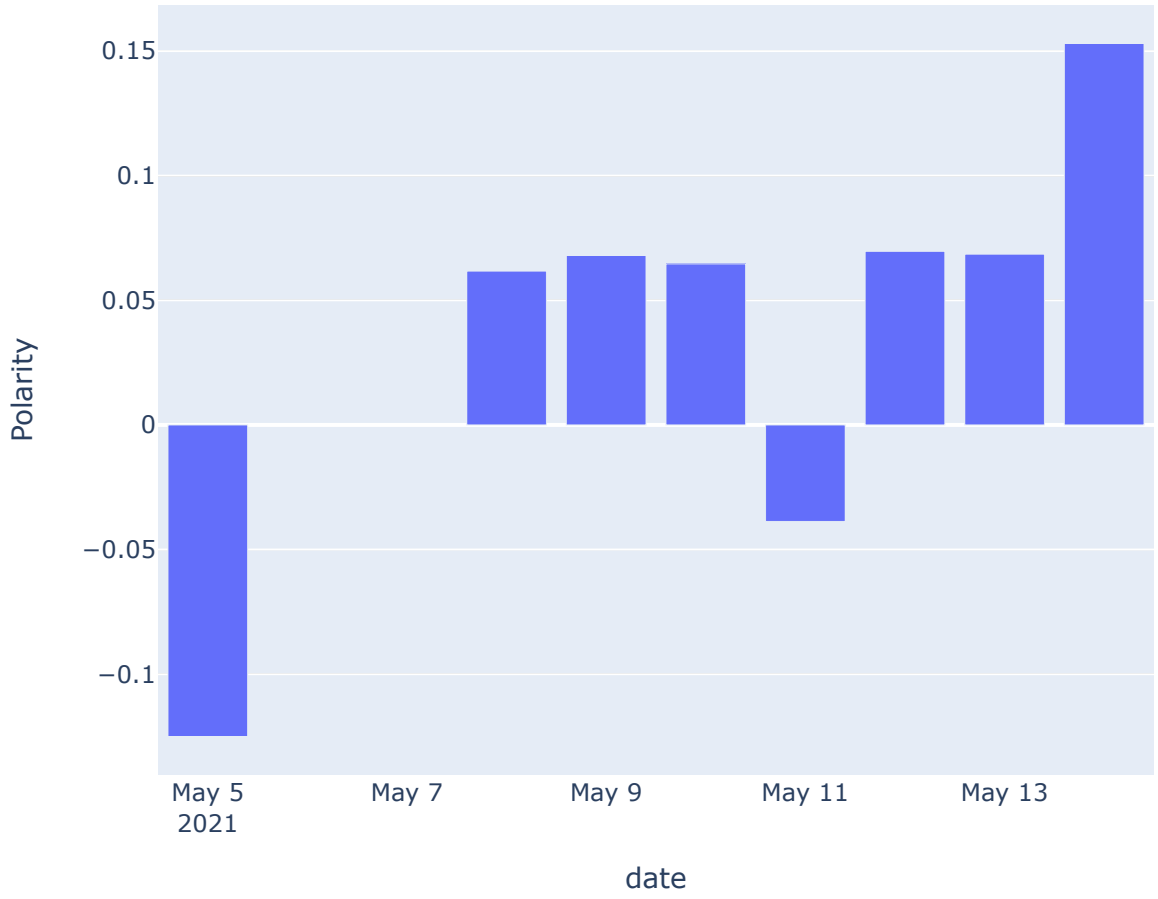
	date	Polarity
0	2021-05-05	-0.125000
1	2021-05-08	0.061720
2	2021-05-09	0.067975
3	2021-05-10	0.064608
4	2021-05-11	-0.038724
5	2021-05-12	0.069634
6	2021-05-13	0.068507
7	2021-05-14	0.153030

In [17]:
```
# export our data into csv file.
```

```
df.to_csv('df_clean_data.csv', index=False)
df.to_csv('df_avg_number_messages.csv', index=False)
```

In [18]:

```
# plot
fig = px.bar(df2, x=df2['date'], y=df2['Polarity'])
fig.show()
```



We can ***conclude*** that:

There were more negative messages on May 5, 2021 and May 11, 2021.

There were more positive messages on other days.