Use NLP to do sentiment analysis on social media data

This project was conducted in the middle of the stress time of Ukraine war, hence I pulled the data on Reddit to see what they were talking and how they reacted with this war:

- Firstly, pulled the data from Reddit with topic is 'Ukraine'.
- Next, extract the sentiment scores of the titles of the posts.
- Plot a histogram of the sentiment scores
- Look at descriptive statistics (mean, median, standard deviation) of the sentiment scores
- Examine the text for some of the highest and lowest sentiment scores
- Finally, I write a short analysis of the results and the process.

Collect social data from Reddit, topic is 'ukraine'

```python
In [10]:   import praw
           import pandas as pd
           import sqlite3
           import pandas as pd
           import credentialsChi
```

```python
In [2]:    reddit = praw.Reddit(client_id=credentialsChi.client_id,
                                client_secret=credentialsChi.client_secret,
                                user_agent=credentialsChi.user_agent)
```

```
Version 7.6.1 of praw is outdated. Version 7.7.0 was released Saturday Februar
y 25, 2023.
```

```python
In [3]:    ukr_subreddit = reddit.subreddit('ukraine').hot(limit=10)
```

```python
In [4]:    for post in ukr_subreddit:
               print(post.title, post.score)
```

6:24 EET; The Sun is Rising Over Kyiv on the 380th Day of the Full-Scale Invas
ion. Today's post is the final entry in our series about Taras Shevchenko, who
died on this day in 1861. 402
Good news. We almost done with DJI Mavic 3T. Need more $2308 and we will buy t
his drone for Ukrainian soldiers in Bakhmut. Hope for your help. All info in t
he comments. 696
For those who worry that standing up to Russia would just provoke Putin and dr
ag the world into war - we only have to look at the history of the 20th centur
y. Nothing is more provocative to a dictator than the weakness of free nation
s. 11121
Many explosions heard in Mariupol! 1033
Farewell ceremony for Hero of Ukraine Dmytro "Da Vinci" Kotsubailo commander o
f 1st Assult Battalion "Da Vinci's wolfs" 1760
On the night of March 10, 2023, partisans burned a combat aircraft near Vladiv
ostok, russia 903
Finland's PM Sanna Marin in Kyiv today pays respects to Dmytro Kotsyubailo "da
Vinci" who died fighting for Bakhmut 1207
The Wagner Group's offensive operation in eastern Bakhmut appears to have ente
red a temporary tactical pause 1443
Armored vehicles. Unknown time/location. Bradleys? 390
Today, on the day of the Ukrainian anthem, President Zelensky published a vide
o showing Ukrainians singing their anthem in different conditions: on the fron
tline, in a bomb shelter, during the Revolution of Dignity as well as during t
he celebration of Independence day. 475

In [5]:
```python
reddit_data = {'title': [],
               'link': [],
               'author': [],
               'n_comments': [],
               'score': [],
               'text': [],
               'id': []}


ukr_subreddit = reddit.subreddit('ukraine').hot(limit=None)

for post in list(ukr_subreddit):
    reddit_data['title'].append(post.title)
    reddit_data['link'].append(post.permalink)
    if post.author is None:
        reddit_data['author'].append('')
    else:
        reddit_data['author'].append(post.author.name)

    reddit_data['n_comments'].append(post.num_comments)
    reddit_data['score'].append(post.score)
    reddit_data['text'].append(post.selftext)
    reddit_data['id'].append(post.id)
```

In [6]:
```python
ukr_df = pd.DataFrame(reddit_data)
ukr_df
```

Out[6]:

	title	link	author	n_comme
0	6:24 EET; The Sun is Rising Over Kyiv on the 3...	/r/ukraine/comments/11neqpy/624_eet_the_sun_is...	duellingislands	
1	Good news. We almost done with DJI Mavic 3T. N...	/r/ukraine/comments/11mraik/good_news_we_almos...	serhiiiam	
2	For those who worry that standing up to Russia...	/r/ukraine/comments/11ngsfp/for_those_who_worr...	pastebluepaste	
3	Many explosions heard in Mariupol!	/r/ukraine/comments/11ns1zo/many_explosions_he...	Kubolomo	
4	Farewell ceremony for Hero of Ukraine Dmytro "...	/r/ukraine/comments/11nnc7m/farewell_ceremony_...	BananaBrumik	
...	...	...	...	
807	It's Always Sunny in Crimea - You can support ...	/r/ukraine/comments/11h52jm/its_always_sunny_i...	Ukraine_Aid_Ops	
808	Destruction of the Russian Tor-M2 air defense ...	/r/ukraine/comments/11h15ce/destruction_of_the...	Kay51995	
809	Video of the downed Su-34.	/r/ukraine/comments/11h2en2/video_of_the_downe...	Geschichtsklitterung	
810	Magyar update from Bakhmut - dated 2 March, po...	/r/ukraine/comments/11gxxmz/magyar_update_from...	RoninSolutions	
811	Zelensky meets with Latvian President in Lviv	/r/ukraine/comments/11h4bwa/zelensky_meets_wit...	KI_official	

812 rows × 7 columns

In [7]:
```python
# Collect comments
reddit_comments_data = {'author': [],
                        'score': [],
                        'text': [],
                        'post_id': []}
ukr_subreddit = reddit.subreddit('ukraine').hot(limit=10)

for post in list(ukr_subreddit):
    post.comments.replace_more(limit=None)
    for comment in post.comments:
        if comment.author is None:
            reddit_comments_data['author'].append('')
        else:
            reddit_comments_data['author'].append(comment.author.name)
        reddit_comments_data['score'].append(comment.score)
        reddit_comments_data['text'].append(comment.body)
        reddit_comments_data['post_id'].append(post.id)
```

In [8]:
```python
comment_df = pd.DataFrame(reddit_comments_data)
comment_df
```

Out[8]:

	author	score	text	post_id
0	AutoModerator	1	\nПривіт u/duellingislands ! During wartime, t...	11neqpy
1	spsteve	17	Unbelievable we are closing in on day 400, whe...	11neqpy
2	Jizzapherina	14	This is still so relevant today:\n\nKeep laugh...	11neqpy
3	sonicboomer46	13	I won't soon forget the inane russian propagan...	11neqpy
4	StevenStephen	7	There can be no wondering why he is so revered...	11neqpy
...	...	...	...	...
277	wunderfullynow	17	This anthem is at once sad, hopeful, defiant a...	11nqsu8
278	freetimerva	11	My fuckin man singing in the trench hit me rig...	11nqsu8
279	Named_User-Name	10	Inspiring!\n\nRussia will never conquer these ...	11nqsu8
280	Rock-it-again	9	Fuck that choked me up pretty good.	11nqsu8
281	Wade8869	7	Slava Ukraini!	11nqsu8

282 rows × 4 columns

In [12]:
```python
# Save data to SQLite3 database
ukr = sqlite3.connect("/Users/kienguyen/Documents/DATA SCIENCE/co_reddit.sqlite
ukr_df.to_sql('posts', ukr, if_exists='replace', index=False)
```

Out[12]: 812

In [13]:
```python
ukr_df_check = pd.read_sql_query('SELECT * FROM posts;', ukr)
ukr.close()
ukr_df_check
```

Out[13]:

	title	link	author	n_comme
0	6:24 EET; The Sun is Rising Over Kyiv on the 3...	/r/ukraine/comments/11neqpy/624_eet_the_sun_is...	duellingislands	
1	Good news. We almost done with DJI Mavic 3T. N...	/r/ukraine/comments/11mraik/good_news_we_almos...	serhiiiam	
2	For those who worry that standing up to Russia...	/r/ukraine/comments/11ngsfp/for_those_who_worr...	pastebluepaste	
3	Many explosions heard in Mariupol!	/r/ukraine/comments/11ns1zo/many_explosions_he...	Kubolomo	
4	Farewell ceremony for Hero of Ukraine Dmytro "...	/r/ukraine/comments/11nnc7m/farewell_ceremony_...	BananaBrumik	
...	...	...	...	
807	It's Always Sunny in Crimea - You can support ...	/r/ukraine/comments/11h52jm/its_always_sunny_i...	Ukraine_Aid_Ops	
808	Destruction of the Russian Tor-M2 air defense ...	/r/ukraine/comments/11h15ce/destruction_of_the...	Kay51995	
809	Video of the downed Su-34.	/r/ukraine/comments/11h2en2/video_of_the_downe...	Geschichtsklitterung	
810	Magyar update from Bakhmut - dated 2 March, po...	/r/ukraine/comments/11gxxmz/magyar_update_from...	RoninSolutions	
811	Zelensky meets with Latvian President in Lviv	/r/ukraine/comments/11h4bwa/zelensky_meets_wit...	KI_official	

812 rows × 7 columns

Sentiment analysis

In [14]:
```python
#Load data
con = sqlite3.connect('/Users/kienguyen/Documents/DATA SCIENCE/co_reddit.sqlite
df = pd.read_sql_query('SELECT * from posts;', con)
con.close()
df
```

Out[14]:

	title	link	author	n_comme
0	6:24 EET; The Sun is Rising Over Kyiv on the 3...	/r/ukraine/comments/11neqpy/624_eet_the_sun_is...	duellingislands	
1	Good news. We almost done with DJI Mavic 3T. N...	/r/ukraine/comments/11mraik/good_news_we_almos...	serhiiiam	
2	For those who worry that standing up to Russia...	/r/ukraine/comments/11ngsfp/for_those_who_worr...	pastebluepaste	
3	Many explosions heard in Mariupol!	/r/ukraine/comments/11ns1zo/many_explosions_he...	Kubolomo	
4	Farewell ceremony for Hero of Ukraine Dmytro "...	/r/ukraine/comments/11nnc7m/farewell_ceremony_...	BananaBrumik	
...	...	...	...	
807	It's Always Sunny in Crimea - You can support ...	/r/ukraine/comments/11h52jm/its_always_sunny_i...	Ukraine_Aid_Ops	
808	Destruction of the Russian Tor-M2 air defense ...	/r/ukraine/comments/11h15ce/destruction_of_the...	Kay51995	
809	Video of the downed Su-34.	/r/ukraine/comments/11h2en2/video_of_the_downe...	Geschichtsklitterung	
810	Magyar update from Bakhmut - dated 2 March, po...	/r/ukraine/comments/11gxxmz/magyar_update_from...	RoninSolutions	
811	Zelensky meets with Latvian President in Lviv	/r/ukraine/comments/11h4bwa/zelensky_meets_wit...	KI_official	

812 rows × 7 columns

```
In [2]:   #extract the sentiment scores of the titles of the posts
```

```
In [15]:  sentiment_df = pd.read_csv('AFINN-en-165.txt', sep='\t', names=['word', 'score'
```

```
In [16]:  sentiment_df
```

Out[16]:

	score
word	
abandon	-2
abandoned	-2
abandons	-2
abducted	-2
abduction	-2
...	...
yucky	-2
yummy	3
zealot	-2
zealots	-2
zealous	2

3382 rows × 1 columns

```
In [17]:  sentiment_dict = sentiment_df.to_dict()['score']
```

```
In [18]:  import numpy as np
          # Keyword method
          title_sentiments = []
          for title in df['title']:
              words = title.lower().split()
              this_titles_sentiments = []
              for w in words:
                  if w in sentiment_dict.keys():
                      this_titles_sentiments.append(sentiment_dict[w])
                  else:
                      this_titles_sentiments.append(0)


              title_sentiments.append(np.mean(this_titles_sentiments))
```

```
In [19]:  df['keyword_sentiment'] = title_sentiments
```

```
In [20]:  df['keyword_sentiment']
```
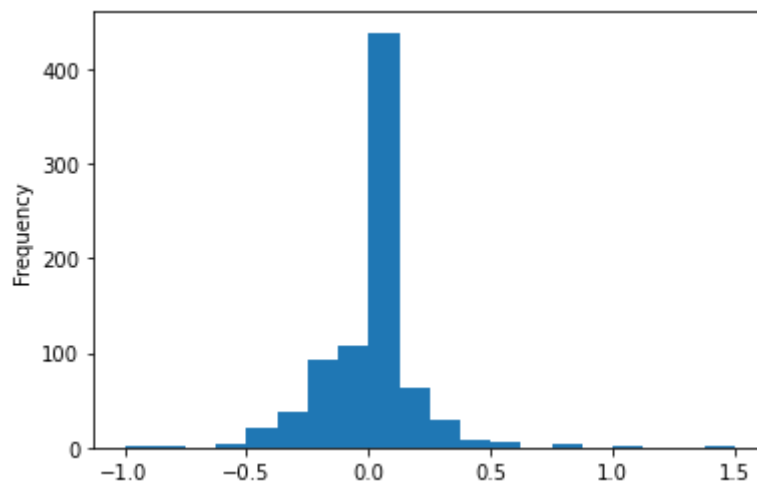
```
Out[20]: 0      -0.085714
         1       0.156250
         2      -0.177778
         3       0.000000
         4       0.111111
                    ...
         807     0.095238
         808    -0.181818
         809     0.000000
         810     0.000000
         811     0.000000
         Name: keyword_sentiment, Length: 812, dtype: float64
```

```
In [21]: df['keyword_sentiment'].plot.hist(bins=20)
```

```
Out[21]: <AxesSubplot:ylabel='Frequency'>
```



```
In [22]: df['keyword_sentiment'].mean()
```

```
Out[22]: -0.014557730048440642
```

```
In [23]: df['keyword_sentiment'].std()
```

```
Out[23]: 0.18239390994780602
```

```
In [24]: df['keyword_sentiment'].median()
```

```
Out[24]: 0.0
```

```
In [13]: #examine the text for some of the highest and lowest sentiment scores
```

```
In [25]: df.sort_values(by='keyword_sentiment')[['title', 'keyword_sentiment']]
```

Out[25]:

	title	keyword_sentiment
422	Defiance over Evil	-1.000000
119	Invaders destroy 400 apartment blocks in Mariu...	-0.769231
652	NATO commander: Russia lost more than 200,000 ...	-0.615385
593	Ukrainian military destroy enemy gun, MLRS on ...	-0.555556
535	Ukraine's Kostyuk dedicates first tennis title...	-0.555556
...	...	...
273	Happy International Women's Day!	0.750000
75	Fight to Win 💛💙🇺🇦	0.750000
196	Ukraine's most committed backer wins a huge el...	0.818182
419	Absolute Courage and Honor	1.000000
357	Nice throw	1.500000

812 rows × 2 columns

In [26]:
```python
df.sort_values(by='keyword_sentiment')['title'].to_list()[:10]
```

Out[26]:
```
['Defiance over Evil',
 'Invaders destroy 400 apartment blocks in Mariupol to hide war crimes — Mayo
r',
 'NATO commander: Russia lost more than 200,000 killed and wounded in the wa
r',
 'Ukrainian military destroy enemy gun, MLRS on Kinburn Spit',
 "Ukraine's Kostyuk dedicates first tennis title to war victims",
 'Negative one more kornet.',
 'Ukraine: How the war is making soil and water toxic',
 "Russia's invasion of Ukraine: Making the perpetrators pay",
 'Ukrainian victims of war with Russia deserve justice, Garland says',
 'Day 375 – The Destruction in This War is Beyond Understanding – War in Ukrai
ne']
```

In [27]:
```python
df.sort_values(by='keyword_sentiment', ascending=False)['title'].to_list()[:10]
```

Out[27]:
```
['Nice throw',
 'Absolute Courage and Honor',
 'Ukraine's most committed backer wins a huge election victory in Estonia',
 'Fight to Win 💛💙🇺🇦',
 'Happy International Women's Day!',
 'Russian engineering at its finest',
 'Fight to win 💛💙🇺🇦 Taras Shevchenko',
 'Ukrainian Rights Group Truth Hounds Wins Norwegian Prize',
 'I-134 form help (family)',
 'Solar panels to donate']
```
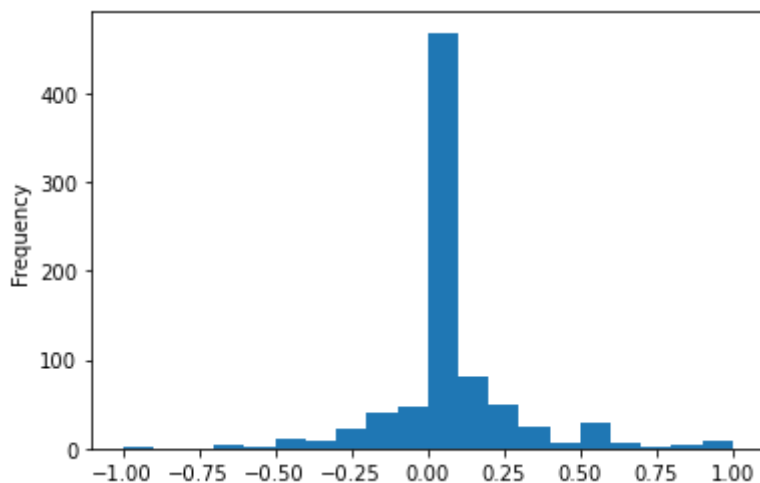
In [29]:
```python
#TextBlob method
from textblob import TextBlob
```

In [30]:
```python
def get_tb_sentiment(text):
    tb = TextBlob(text)
    return tb.sentiment[0]
```

```python
df['tb_sentiment'] = df['title'].apply(get_tb_sentiment)
```

In [31]:
```python
df['tb_sentiment'].plot.hist(bins=20)
```

Out[31]:
```
<AxesSubplot:ylabel='Frequency'>
```


In [32]:
```python
df['tb_sentiment'].mean()
```

Out[32]:
```
0.05859302194623933
```

In [33]:
```python
df['tb_sentiment'].std()
```

Out[33]:
```
0.21300704107920854
```

In [34]:
```python
df['tb_sentiment'].median()
```

Out[34]:
```
0.0
```

In [25]:
```python
#VADER method
```

In [35]:
```python
import nltk
nltk.download('vader_lexicon')
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     /Users/kienguyen/nltk_data...
```

In [36]:
```python
analyzer = SentimentIntensityAnalyzer()
df['vd_sentiment'] = df['title'].apply(lambda title: analyzer.polarity_scores(t
df.head()
```

Out[36]:

|	title	link	author	n_comments	sco
0	6:24 EET; The Sun is Rising Over Kyiv on the 3...	/r/ukraine/comments/11neqpy/624_eet_the_sun_is...	duellingislands	9	4
1	Good news. We almost done with DJI Mavic 3T. N...	/r/ukraine/comments/11mraik/good_news_we_almos...	serhiiiam	23	7
2	For those who worry that standing up to Russia...	/r/ukraine/comments/11ngsfp/for_those_who_worr...	pastebluepaste	262	11
3	Many explosions heard in Mariupol!	/r/ukraine/comments/11ns1zo/many_explosions_he...	Kubolomo	104	10
4	Farewell ceremony for Hero of Ukraine Dmytro "...	/r/ukraine/comments/11nnc7m/farewell_ceremony_...	BananaBrumik	46	17

In [37]:
```python
df['vd_compound'] = df['vd_sentiment'].apply(lambda score_dict: score_dict['con
df['sentiment_type']=''
df.loc[df.vd_compound>0,'sentiment_type']='POSITIVE'
df.loc[df.vd_compound==0,'sentiment_type']='NEUTRAL'
df.loc[df.vd_compound<0,'sentiment_type']='NEGATIVE'
```

In [38]:
```python
df.sentiment_type.value_counts().plot(kind='bar',title="sentiment analysis")
```
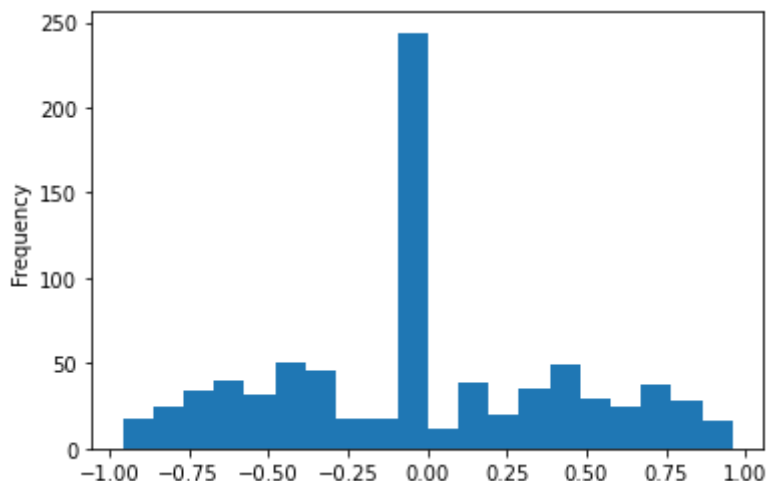
Out[38]:
```
<AxesSubplot:title={'center':'sentiment analysis'}>
```
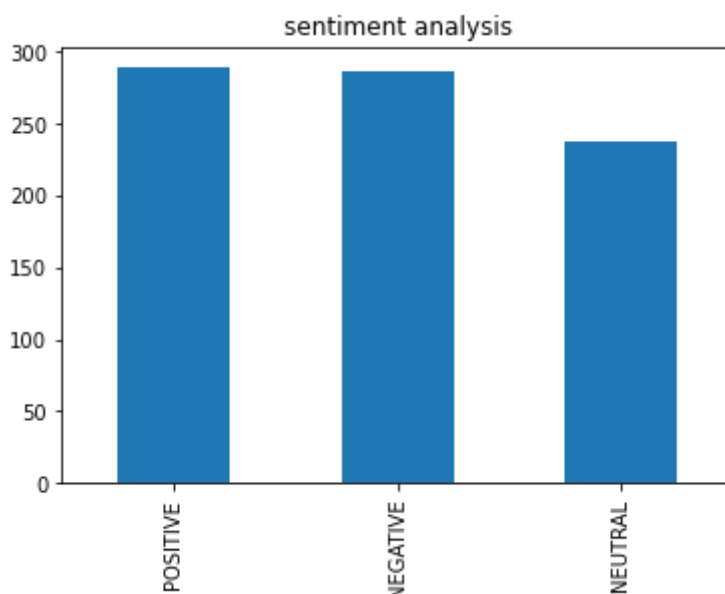
In [39]:
```python
df['vd_compound'].plot.hist(bins=20)
```

Out[39]:
```
<AxesSubplot:ylabel='Frequency'>
```


In [43]:
```python
df['vd_compound'].mean()
```

Out[43]:
```
-0.011711330049261094
```

In [41]:
```python
df['vd_compound'].std()
```

Out[41]:
```
0.46201720076734004
```

In [42]:
```python
df['vd_compound'].median()
```

Out[42]:
```
0.0
```

Summary

Firstly, I tried to extract the sentiment scores of the titles of the posts by using keyword method, and the scores are mainly around 0, however the average score is -0.015. Next, I tried more with Textblob and Vader to check the sentiment scores between these methods. It's different among these 3 methods, the Textblob method gives us a positive average scores, while the other two resulted in negative mean scores, the Vader method and Keyword socres are almost -0.012 vs -0.02. However, the standard deviation is nearly the same between Keyword and Textblob methods, around 0.2.

In general, the content about "Ukraine" on Reddit seems negative according to the Keyword and Vader methods.

According to the Keyword method, the content of the posts having highest sentiment score is mostly to encourage Ukraine and the posts having lowest scores is mostly to against the war.

In my opinion, these findings can help journalists or communication companies understand about the people who care about the status of Ukraine to write and issue articles with

proper topics and words.