TASK 4:

In [3]:
```
import pandas as pd
from textblob import TextBlob
import matplotlib.pyplot as plt
```

In [4]:
```
data= pd.read_csv('twitter_training.csv')
```

In [5]:
```
data.head()
```

Out[5]:

	target	word	label	tweet
0	2401	Borderlands	Positive	im getting on borderlands and i will murder yo...
1	2401	Borderlands	Positive	I am coming to the borders and I will kill you...
2	2401	Borderlands	Positive	im getting on borderlands and i will kill you ...
3	2401	Borderlands	Positive	im coming on borderlands and i will murder you...
4	2401	Borderlands	Positive	im getting on borderlands 2 and i will murder ...

In [6]:
```
data.tail()
```

Out[6]:

	target	word	label	tweet
74677	9200	Nvidia	Positive	Just realized that the Windows partition of my...
74678	9200	Nvidia	Positive	Just realized that my Mac window partition is ...
74679	9200	Nvidia	Positive	Just realized the windows partition of my Mac ...
74680	9200	Nvidia	Positive	Just realized between the windows partition of...
74681	9200	Nvidia	Positive	Just like the windows partition of my Mac is l...

In [7]:
```
data.columns
```

Out[7]:
```
Index(['target', 'word', 'label', 'tweet'], dtype='object')
```

In [8]:
```
data.shape
```

Out[8]:
```
(74682, 4)
```

In [9]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 74682 entries, 0 to 74681
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   target  74682 non-null  int64 
 1   word    74682 non-null  object
 2   label   74682 non-null  object
 3   tweet   73996 non-null  object
dtypes: int64(1), object(3)
memory usage: 2.3+ MB
```

In [10]:
```
col_names=['ID','Entity','Sentiments','Contest']
df=pd.read_csv('twitter_training.csv', names=col_names)
```

In [11]:
```
df.head()
```

Out[11]:

	ID	Entity	Sentiments	Contest
0	target	word	label	tweet
1	2401	Borderlands	Positive	im getting on borderlands and i will murder yo...
2	2401	Borderlands	Positive	I am coming to the borders and I will kill you...
3	2401	Borderlands	Positive	im getting on borderlands and i will kill you ...
4	2401	Borderlands	Positive	im coming on borderlands and i will murder you...

In [12]:
```
df.shape
```

Out[12]:
```
(74683, 4)
```

In [14]:
```
df.describe
```

Out[14]:
```
<bound method NDFrame.describe of            ID       Entity Sentiments  \
0      target         word      label
1        2401  Borderlands   Positive
2        2401  Borderlands   Positive
3        2401  Borderlands   Positive
4        2401  Borderlands   Positive
...       ...          ...        ...
74678    9200       Nvidia   Positive
74679    9200       Nvidia   Positive
74680    9200       Nvidia   Positive
74681    9200       Nvidia   Positive
74682    9200       Nvidia   Positive

                                          Contest
0                                           tweet
1      im getting on borderlands and i will murder yo...
2      I am coming to the borders and I will kill you...
3      im getting on borderlands and i will kill you ...
4      im coming on borderlands and i will murder you...
...                                              ...
74678  Just realized that the Windows partition of my...
74679  Just realized that my Mac window partition is ...
74680  Just realized the windows partition of my Mac ...
74681  Just realized between the windows partition of...
74682  Just like the windows partition of my Mac is l...

[74683 rows x 4 columns]>
```

In [15]:
```
df.describe()
```

Out[15]:

	ID	Entity	Sentiments	Contest
count	74683	74683	74683	73997
unique	12448	33	5	69490
top	5203	Microsoft	Negative	At the same time, despite the fact that there ...
freq	6	2400	22542	172

In [16]:
```
df.isnull().sum()
```

Out[16]:
```
ID              0
Entity          0
Sentiments      0
Contest       686
dtype: int64
```

In [17]:
```
df.dropna(axis=0,inplace=True)
```

In [18]:
```
df.isnull().sum()
```

Out[18]:
```
ID            0
Entity        0
Sentiments    0
Contest       0
dtype: int64
```

In [19]:
```
df.duplicated().sum()
```

Out[19]:
```
np.int64(2341)
```

In [21]:
```
df.drop_duplicates(inplace=True)
df.duplicated().sum()
```

Out[21]:
```
np.int64(0)
```

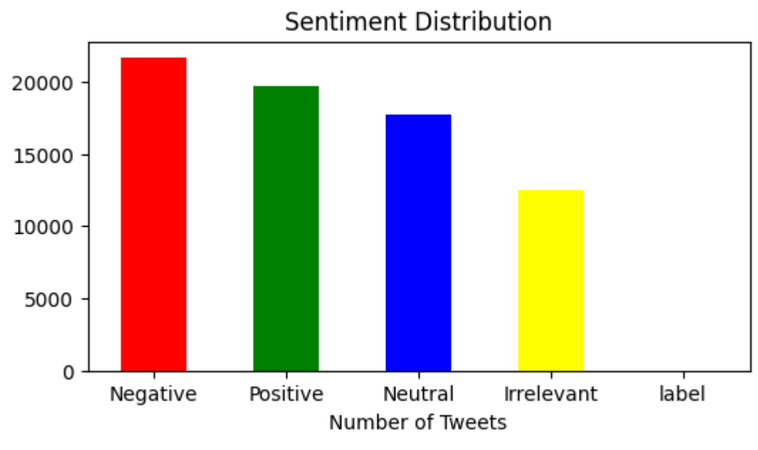
In [23]:
```
df.shape
```

Out[23]:
```
(71656, 4)
```

In [24]:
```
sentiment_counts=df['Sentiments'].value_counts()
sentiment_counts
```

Out[24]:
```
Sentiments
Negative      21698
Positive      19713
Neutral       17707
Irrelevant    12537
label             1
Name: count, dtype: int64
```

In [25]:
```
plt.figure(figsize=(6,3))
sentiment_counts.plot(kind='bar',color=['red','green','blue','yellow'])
plt.title('Sentiment Distribution')
plt.xlabel('Number of Tweets')
plt.xticks(rotation=0)
plt.show()
```


Sentiment Distribution
20000
15000
10000
5000
0
Negative
Positive
Neutral
Irrelevant
label
Number of Tweets

In [ ]: