
0.0.1 Question 1d

There are many ways we could choose to read tweets. Why might someone be interested in doing data analysis on tweets? Name a kind of person or institution which might be interested in this kind of analysis. Then, give two reasons why a data analysis of tweets might be interesting or useful for them. Answer in 2-3 sentences.

Social media researcher. Reasons: (1)The tweets of public figures may infulence public opinion direction. (2)Data analysis of tweets can be used to mine the relationship between tweet sending time and page views, the relationship between word count and page views, etc.

0.0.2 Question 2e

What might we want to investigate further? Write a few sentences below.

Which Twitter platforms users like to use, and the active time of users on various Twitter platforms.etc.

0.0.3 Question 2f

We just looked at the top 5 most commonly used devices for each user. However, we used the number of tweets as a measure, when it might be better to compare these distributions by comparing *proportions* of tweets. Why might proportions of tweets be better measures than numbers of tweets?

Proportion can be more intuitive to see the distribution of data, as well as the relationship between the number of subcategories and the total amount.

0.0.4 Question 3b

Compare Cristiano's distribution with those of AOC and Elon Musk. In particular, compare the distributions before and after Hour 6. What differences did you notice? What might be a possible cause of that? Do the data plotted above seem reasonable?

All these three distributions are skewed right. Before Hour 6, the lines of AOC and Elonmusk present an overall downward trend, while the line of Cristiano presents an overall upward trend. All the three lines reach their peak at about Hour 17.

0.0.5 Question 4a

Please score the sentiment of one of the following words, using your own personal interpretation. No code is required for this question!

- police
- order
- Democrat
- Republican
- gun
- dog
- technology
- TikTok
- security
- face-mask
- science
- climate change
- vaccine

What score did you give it and why? Can you think of a situation in which this word would carry the opposite sentiment to the one you've just assigned?

climate change: I will gice it -1.0 since climate is a serious problem.

opposite situation: The United Nations issues a strict carbon dioxide emission regulation, which can receive positive response from all countries and has a positive effect on regulating climate change

---

0.0.6 Question 4g

When grouping by mentions and aggregating the polarity of the tweets, what aggregation function should we use? What might be one drawback of using the mean?

median function. The drawback of using the mean is that they are susceptible to extreme values.

0.0.7 Question 5a

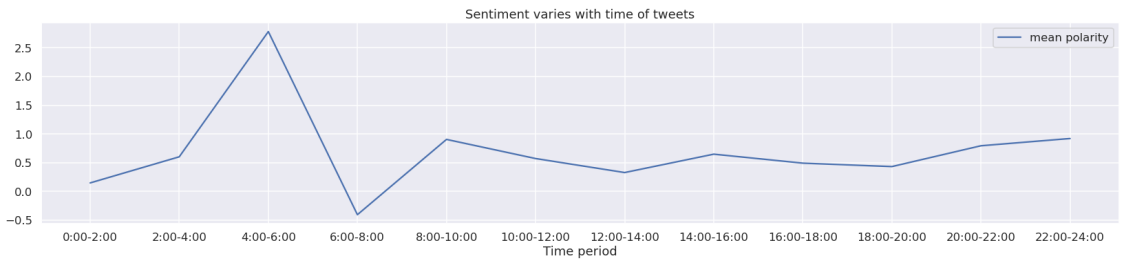
Use this space to put your EDA code.

```
In [44]: def func(x):
             return x // 2
         tweets['AOC']['seg'] = tweets['AOC']['converted_hour'].apply(func)  #calculate the int part of
         res = tweets['AOC'].groupby('seg')[['polarity']].agg(np.mean)  #group by the 'seg' column and
         res = res.rename(columns = {'polarity':'mean polarity'})
         res = res.set_index(pd.Index(['0:00-2:00','2:00-4:00','4:00-6:00','6:00-8:00','8:00-10:00','10
         print(res)

         plt.figure(figsize=(25,5))
         sns.lineplot(data = res)
         plt.xlabel('Time period')
         plt.title('Sentiment varies with time of tweets')
```

```
             mean polarity
0:00-2:00         0.144643
2:00-4:00         0.600000
4:00-6:00         2.783333
6:00-8:00        -0.409375
8:00-10:00        0.903209
10:00-12:00       0.569773
12:00-14:00       0.325342
14:00-16:00       0.646032
16:00-18:00       0.489310
18:00-20:00       0.429443
20:00-22:00       0.792629
22:00-24:00       0.918868
```

```
Out[44]: Text(0.5, 1.0, 'Sentiment varies with time of tweets')
```



0.0.8 Question 5b

Use this space to put your EDA description.

The purpose of the above EDA is to explore how users' sentiment varies with time of posting tweets.

In order to achieve it, I choose 'converted_hour' column as the base time. To better group time together, I calculate the int part of ('converted_hour'/2) and store in a new column 'seg' and take the time into 12 groups.Then I counts the mean of polarity of each group and obtain the average of twitter sentiment each time period.

According to the result plot,we can conclude that the mean polarity reaches the peak during 4:00-6:00am and reaches the trought during 6:00-8:00am, which shows that sentiments are highest in the 4:00-6:00am, lowest in the 6:00-8:00am and remain flat during 12:00-24:00.