
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 that 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.

Traders and financial institutions might be interested in analyzing tweets to understand how the volume or content of tweets can impact stock performance. For instance, influential figures like Elon Musk have been known to significantly sway stock prices with their tweets.

0.0.2 Question 2e

Given the plot above, what might we want to investigate during EDA? Name some possible questions you may have about the dataset in light of the information shown in the plot.

- We can investigating device popularity across different continents can provide a richer understanding of regional preferences and technological trends (Cristiano vs Elon Musk).
- If we want to find out devices used we might want to analyze what kind of devices are used to access specific platforms or websites like WhoSay or MobioINsider.com.

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 (i.e., what percentage of all tweets for a user were published from each device). Why might the proportions of tweets be better measures than the number of tweets?

To see the distributions among different groups and to standardize the data and also because it would not be accurate to compare different groups or categories as raw counts can be misleading if the groups have different sizes.

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?

Hint: If you are not familiar with who Cristiano, AOC, and Elon Musk are, it may be helpful to Google information about these people, their occupations, and where they live.

Cristiano is on GMT time zone whereas AOC and Elon Musk lives on EST/PST time zones. So it make sense for Elon Musk and AOC to have similar distributions since they are most likely around 1 hour different and have peaks during Cristiano's night time/troughs.

0.0.5 Question 4a

Using your own personal interpretation, please score the sentiment of one of the following words using the VADER scale (-4 means the word is extremely negative. +4 means the word is extremely positive). No code is required for this question!

- order
- dog
- cat
- technology
- TikTok
- security
- science
- climate change

What score did you give it and why? Can you describe a situation where this word would carry the opposite sentiment to the one you've just assigned? If not, explain why.

dog: +2.5, people often shares about their pet dogs or dogs they spotted from parks etc. Some people also may share about their pet dogs passing etc and may have different sentiment values.

0.0.6 Question 4g

In `q4f` above, we aggregated the polarity of the tweets by computing the mean sentiment score of tweets mentioning each user. What are some drawbacks of the decision to use the mean as an aggregation function? What other aggregation function(s) might be more appropriate than the mean?

Mean is very sensitive to outliers. We can use median if we have skewed distributions or many outliers that might shift the mean significantly.

0.0.7 Question 5a

Use this space to put your EDA code.

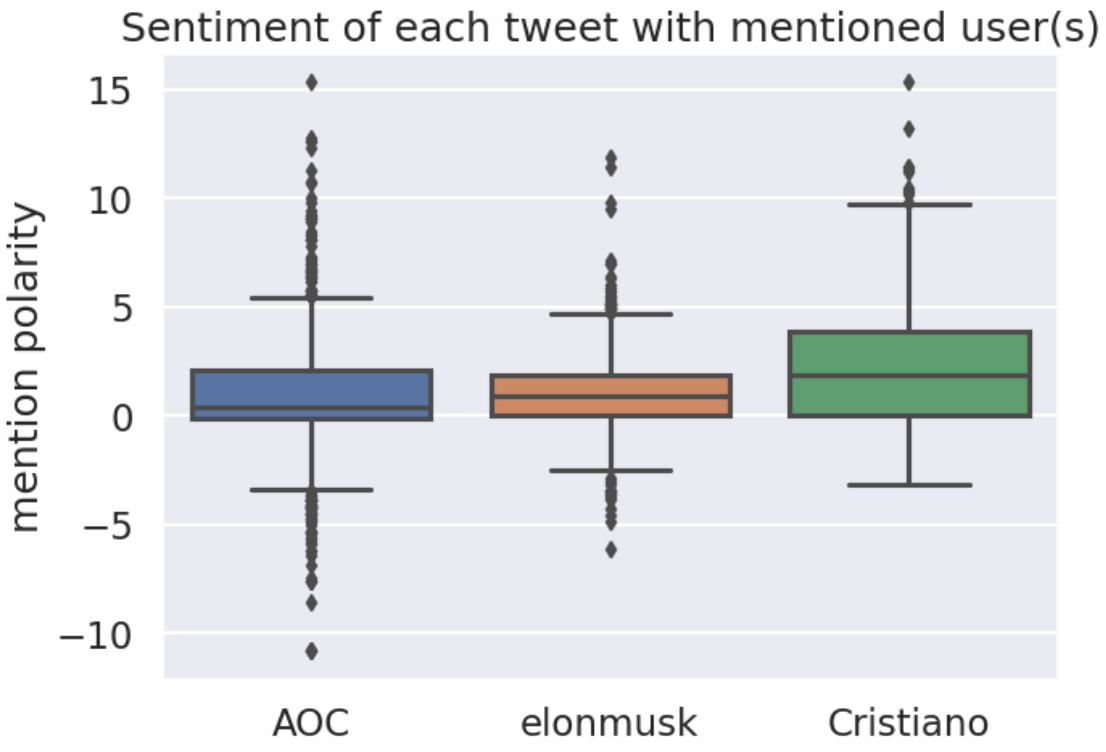
In [183]:
```
aoc_mention_polarity = mention_polarity(tweets["AOC"],mentions["AOC"]).sort_values(ascending=
elon_mention_polarity = mention_polarity(tweets["elonmusk"],mentions["elonmusk"]).sort_values
cris_mention_polarity = mention_polarity(tweets["Cristiano"],mentions["Cristiano"]).sort_valu

mentions_polarities_df = pd.DataFrame({"AOC": aoc_mention_polarity, "elonmusk": elon_mention_

sns.boxplot(data=mentions_polarities_df)

plt.ylabel('mention polarity')
plt.title('Sentiment of each tweet with mentioned user(s)')
```

Out[183]:
```
Text(0.5, 1.0, 'Sentiment of each tweet with mentioned user(s)')
```



0.0.8 Question 5b

Use this space to put your EDA description.

It would be useful to find out their public perception, to gauge the sentiment of tweets that mention these users can help understand the general sentiment or mood of the public towards them. Are the sentiments predominantly positive, negative, or neutral? Here we can see that Cristiano has higher mean sentiment as compared to AOC and Elonmusk. AOC's sentiment scores that exhibit extreme values/outliers could imply several things such as her polarizing figure or even general twitter dynamics.