
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.

Commercial advertisement companies, political campaign poll collectors, research institutes of social science can all be interested in analyzing tweet data.

For commercial advertisement companies, they wish to know how many people viewed their ads or clicked the tweet/link. They broadcast the ads through collaboration with celebrities or through the platform. By analyzing tweet analytics, they can understand 1) what kind of audience are attracted, through which channel; 2) whether the ads have turned into profitability; 3) whether the ads are worth the investment and garnering attention.

0.0.2 Question 2e

What might we want to investigate further based on the plot in 2d above? Write a few sentences below.

1. I would like to understand whether there's any time-dependent trend in the tweets, e.g. whether the tweets were posted in the day/night, workdays vs weekends, is the frequency increased or decreased over the years or any specific event, etc.
2. I would like to analyze any speech pattern of the tweets, see if I can differentiate tweets posted by the owner or by the managing company. E.g. I can use Cristino's tweets as a training set to do supervised learning, then use the model to predict AOC and elon musk's tweets.

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?

What we want to know from this information is how often does each user use a device to tweet. Hence proportions of tweets by device would be a good way to show us the percentage each device is used. It provides a wholistic picture with a clear comparison. Number of tweets, if not knowing the total number of tweets, will not be informative enough. It's nothing more than a number.

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?

Cristiano's tweets were posted mostly in the daytime. From 12-6am there were not many tweets. Elon Musk and AOC have a lot of tweets even during the normal sleeping hours.

One possible cause is that Cristiano and AOC sent most of the tweets themselves, or their management companies were in the same timezone. Elon Musk likely had some management companies tweeting for him throughout the clock. Or there's also likelihood that Elon Musk doesn't really sleep in the night time. AOC seem to be dormant 6-12am, which might be her sleeping schedule and the difference between AOC and Cristiano is likely due to timezone. Elon Musk has no stop whatsoever.

The line plot is very informative in revealing a summed tweeting pattern throughout the clock for the celebrities. It also brings an interesting possibility that sticking to the same tweeting platform may not mean that the owner tweeted everything himself. However, due to the timezone difference, the data may not be 100% fair to compare. It's worth additional analytics.

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?

- police: 2. I think police generally represents justice and security. But there might be occasions where police is related to bully, dictatorship or misuse of violent forces.
- order: 3. I think order if generally representing a stabilized, non-violent environment. But when it comes to the council, congress or courts, order can be shouted in a yelling tone and appear more mandating and negative.
- Democrat: 0. I think it's a party of certain political advocates. It's neither positive nor negative in my opinion. But in cases like a republican gathering, they might shout out Democrat in a negative tone.
- Republican: 0. I think it's a party of certain political advocates. It's neither positive nor negative in my opinion. But in cases like a Democrat gathering, they might shout out Democrat in a negative tone.
- gun: -3. I think it's a tool of extreme violence. People created this tool to kill. Only in cases when certain people use guns to protect themselves or their motherland would this word be given a little positive sentiment. But still if the guns are not created at the first place there won't be associated violence.
- dog: 3. I think dogs are generally raised and trained to be friends of people. There can be cases where people curse with "dog" in their wording, making it a little negative.
- technology: 3. I think technology is generally positive, that represents human intellectual and engineering advancement. But when technology is associated with violence / or is against human rights (like surveillence technology), it's negative.

- TikTok: 1. I think TikTok is neutral positive. It's a platform and technology advancement. It can also be related to spying or addiction, which appears negative.
- security: 3. I think security is almost always a positive word, like home security, job security, etc. But in occassions where someone calls "security!", it might mean a bunch of armed guards coming out to wrestle and threaten. This scene could be a little negative.
- face-mask: 0. I think it's a very neutral word. But it can be negative when all human beings are forced to wear face masks every day.
- science: 3.5. I think it's a very positive word. When it comes to science fraud it might be negative.
- climate change: 0. I think it's neutral. But it seems always related to bad issues such as extreme weathers, droughts, etc, where the word appears negative.
- vaccine: 3.5. I think it's a very positive word. When it's related to medical accidents the word can be negative.

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?

```
In [49]: aoc_mention_polarity.groupby('mentions').agg(np.mean)
```

```
Out[49]: mentions
         1010wins          2.95
         1mind_the_gap     4.70
         350              -0.10
         4lisaguerrero     4.80
         7                 1.10
                           …
         zellnor4ny        0.80
         zephyrteachout   -4.70
         zerlinama         0.60
         zerlinamaxwell    0.00
         zeynep            0.00
         Name: polarity, Length: 1182, dtype: float64
```

I used the following function: ".groupby('mentions').agg(np.mean)"

One issue is the mean represents average of all the scores. It doesn't show distribution, as the scores might be polarized and have two peaks. But their mean could sit inbetween. I think a histogram like patter is better.

0.0.7 Question 5a

Use this space to put your EDA code.

```
In [75]: # perform your text analysis here
         hashtags_re = '\#(\w+)'

         def extract_hashtags(full_texts):
             hashtags = full_texts.str.findall(hashtags_re).explode().dropna().str.lower().to_frame().r
             return hashtags[["hashtags"]]

         display(extract_hashtags(tweets["AOC"]["full_text"]).head())

         hashtags = {handle: extract_hashtags(df["full_text"]) for handle, df in tweets.items()}
         horiz_concat_df(hashtags).head()
```

```
                            hashtags
id
1355363792545263617          covid19
1354577938767818764           inners
1354211627940384768       freemariana
1354211627940384768        abolishice
1352095493950754817  huntspointstrike
```

Out[75]:

	AOC	Cristiano	elonmusk
	hashtags	hashtags	hashtags
0	covid19	finoallafine	spacex
1	inners	finoallafine	spacexstarship
2	freemariana	finoallafine	dragon
3	abolishice	finoallafine	crewdragon
4	huntspointstrike	finoallafine	starship

0.0.8 Question 5b

Use this space to put your EDA description.

I would like to expand from using regular expressions to find out mentions, to explore hashtags.

I first created a regular expression `hashtags_re` to represent all strings that start with `#` and extract only the alphanumerical strings that follow the `#`.

I then modified the original function to extract `mentions` to now extract all the `hashtags`. A showcase of the function essentially provides me a Series with `id` as the index and `hashtags` as the column.

By using the horizontal concatenation function, I joined and displayed the three celebrities tweet hashtags