
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.

A marketing company can benefit significantly from analyzing tweets for audience insights and sentiment analysis. Studying the tweets provides the company with demographic information (age, gender, location, interests, etc.) of the consumers of their brand. The company marketers can gauge public sentiment towards their brand by analyzing the sentiment expressed in tweets. For example, negative sentiment may signal issues that need attention. What is more, tracking the usage of brand-specific hashtags or keywords in tweets allows marketers to measure their reach and impact.

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 observe that most of AOC's and Musk's tweets originate from iPhones, while Cristiano has used a variety of devices for tweeting. This prompts us to explore the reasons behind Cristiano's diverse device usage. Additionally, it's worth investigating that Elon Musk employed a unique approach by using the Twitter web app to compose his tweets during EDA.

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?

Comparing the proportions of tweets is preferable because it normalizes the data, enabling fair comparisons across individuals who use Twitter at different frequencies. This approach offers insights into device preferences and usage patterns, making it effective for identifying changes in behavior.

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 barely tweets before Hour 6, but his tweets increase significantly after that hour. Both AOC and Elon Musk tweets actively before Hour 6, but their tweets decrease after that hour, with AOC's falls down to almost no tweets. All three of them reach the peak of their tweets around Hour 16 and 17. This difference in trend can be a cause of the time differences between their locations: Cristiano lives in Europe, whereas AOC and Elon lives in the US.

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: +3 I assigned a high positive score because "dog" is often associated with affection, companionship, and happiness. However, in situations involving fear or aggression, it could carry a negative sentiment. For example, a tweet on neighbor's dog barking all night will be associated with negative sentiment.

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?

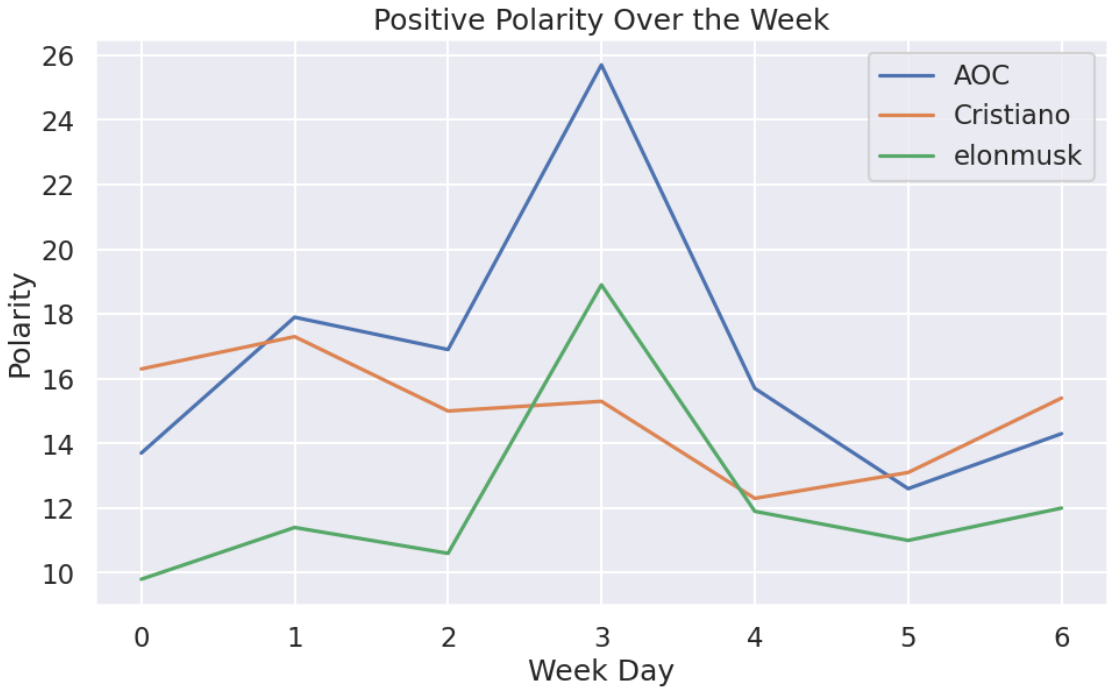
Using the mean as an aggregation function for sentiment scores in tweets can be problematic because it's sensitive to outliers, which can potentially skew the representation of sentiment. For example, if a user posts extremely negative contents, it may drop the the overall polarity. A more appropriate aggregation method, such as the median, can provide a more robust representation of central sentiment tendency while minimizing the influence of extreme values.

0.0.7 Question 5a

Use this space to put your EDA code.

```
In [48]: def highest_polarity(df):
             df["week_day"] = df["converted_time"].apply(lambda x: x.weekday())
             return df[["week_day", "polarity"]].groupby("week_day").agg("max")
         hp = {handle: highest_polarity(df) for handle, df in tweets.items()}

         make_line_plot(hp, "week_day", "polarity",
                        title="Positive Polarity Over the Week",
                        xlabel="Week Day", ylabel="Polarity")
```



0.0.8 Question 5b

Use this space to put your EDA description.

I am investigating the most postive time during a week for each of the three users. For both AOC and Elon Musk, the most positive week day is '3', which is Thursday because the week day index starts from 0. For Cristiano, the most positive day of the week is Tuesday. We notice that Cristiano's polarity remains relatively consistent throughout the week, while the two other users have a much higher polarity on their most positive day (probably due to a combination of anticipation for the weekend and strategic timing for positive content releases).