
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.

- Twitter Analytics gives individuals basic details of their followers and it allows to to view audience insights for all Twitter users, their followers and their organic audience. Additionally, analyzing Tweet dadas give companies and investors useful insights about thier services and products quality and specifically, they will be able to understand what's working well and what's not, so they improve areas needs improvement based on their audience expectations. Text Analysis Pedagogy Institute is one of the prominent institutes that performs text analysis.

0.0.2 Question 2e

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

Although most of the tweets from these three famous individuals are from Twitter for iPhone, but unlike Elon Musk and AOC, Cristiano's Tweets are from different web clients that make us curious about the accuracy of his Tweets. For instance, WhoSay and MobioINsider are both marketing web clients that may abuse Cristiano's Tweets based on thier own benifits. So, these areas are diffenitely need to be invistigated.

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?

Using proportions help us compare relative qualnitiies of groups and measure changes in quantities (in this case the proportions of Tweets) for different individuals in our data. For instance, if Cristiano's Tweeter for iPhone were more than in number compared to Elon Musk we'd have overestimated that Cristiano is much more prominent iPhone user, although Elon Musk is more prominent iPhone user because Cristiano used several other devices which basically means Cristiano is not a prominent iPhone user. So, using proportions of tweets are much more helpful to understand such implications compared to numbers as a measurement.

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?

If we start from 0-5 or 6 hours, we see that Cristiano is literally not posting anything compared to AOC and Elon Musk. It means that Cristiano posts mostly after 6 in the morning, so his tweet's distribution is focused from 6am - 10pm or so. However, Elon Musk and AOC are more active before 6am but their distributions declines after 8pm. Another notion is that, perhaps, AOC may have a different sleep time schedualed comapred to Cristiano, which make sense because a soccer player exercises in the morning while a politician maybe sleeping at that time due to working late at nights. Elon Musk is distribution goes down from 8am - 02pm, perhaps, that's his working and meeting time, so he might posts very important or time sensetive issues or during weekends at that specific time.

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: -0.1 although police are to protect us, but police brutality in recent years had negative impact on overall people's thoughts and lives. Order: 0.1 one reason is that based in my own experience I have always thought order means discipline. Democrat: -0.2 because their descision made almost 35 millions of Afghans suffer from the Taliban regime. Republican: -0.3 because they spread recism in the country while democracy and equality is the base of this country Gun: -0.5 the worse thing ever in the United States Dog: -0.9 I love dogs for many reasons Technology: 0.5 technology is good if we have the knowledge on how to utilize it TikTok: -0.2 has a negative impact on overall society Security: 0.8 we need peace to develop and security is one of it's key component Face-mask: -0.1: I had a bad experience during covid, so I don't like face masks at all Science: 0.9 science gives us a reason to live in this wold climate change: -0.4 climate change is like cancer that never goes away Vaccine: -0.2 only because it reminds of deaths during covid

---

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?

One of the other choices would the median aggregate function as we learned in our discussion section, mean is not always the best choice when dealing with small sample sizes and there is an outlier in the data.

0.0.7 Question 5a

Use this space to put your EDA code.

In [ ]:

In [94]:
```
hashtag_re = r"\#([\w]*)"

def extract_hash(text, hashtag):
    hashtag = text.str.lower().str.extractall(hashtag)
    hashtag = hashtag.rename(columns={0: "hashtag"}).reset_index(level=1)
    return hashtag[["hashtag"]]


def hashtag_polarity(df, hash_df, tag):
    merged = pd.merge(hash_df, df["polarity"], left_index=True, right_index=True)
    if(tag==1):
        return merged.dropna().groupby("hashtag").mean()["polarity"]
    else:
        return merged.dropna().groupby("hashtag").median()["polarity"]

def compare(name):
    return pd.DataFrame({"mean": hashtag_polarity(tweets[name], hashtag[name], 1),
                         "median": hashtag_polarity(tweets[name], hashtag[name], 2)
                        })
hashtag_mean = {handle: extract_hash(df["full_text"], hashtag_re) for handle, df in tweets.ite
#horiz_concat_df(mentions).head()

aoc_comparison = compare("AOC").sort_values(by="median", ascending=True)
elonmusk_comparison = compare("elonmusk").sort_values(by="median", ascending=True)
cris_comparison = compare("Cristiano").sort_values(by="median", ascending=True)

print("AOC:")
print(aoc_comparison, "\n")

print("Elon Musk:")
print(elonmusk_comparison, "\n")

print("Cristiano:")
print(cris_comparison)
```

```
AOC:
                   mean  median
hashtag
rayshardbrooks    -15.20  -15.20
```

```
cancelstudentdebt  -6.10   -6.10
cantpaymay         -5.20   -5.20
m4a                -4.90   -4.90
layleenpolanco     -4.60   -4.60
…                  …     …
alopecia            7.50    7.50
happyhanukkah       8.75    8.75
mo1                 8.80    8.80
juneteenth          9.40    9.40
il03               13.10   13.10

[114 rows x 2 columns]

Elon Musk:
                        mean  median
hashtag
justiceforgeorge      -1.200   -1.20
cancelnewsnetwork      0.000    0.00
spacexstarship         0.000    0.00
cybertruck             0.000    0.00
dragon                 0.500    0.00
maythefourthbewithyou  0.000    0.00
tesla                  0.000    0.00
crewdragon             0.050    0.05
spacex                 0.750    0.75
mdtismobilizing        1.200    1.20
launchamerica          1.875    1.50
crew1                  1.500    1.50
nasa                   1.500    1.50
veteransday            1.500    1.50
powerwall              2.000    2.00
sn8                    2.900    2.90
starship               2.900    2.90

Cristiano:
                   mean  median
hashtag
prayers4paris     -6.70   -6.70
nodoubts          -3.60   -3.60
cr7crunchfitness  -2.40   -2.40
gym               -2.40   -2.40
ocnn              -2.20   -2.20
…                 …     …
vivaportugal      10.20   10.20
abbott            10.30   10.30
happybirthdaycr7  10.35   10.35
globesoccer       11.80   11.80
goldenfoot2020    15.40   15.40

[419 rows x 2 columns]
```

0.0.8 Question 5b

Use this space to put your EDA description.

This was an excellent experiment. In this EDA, I had two goals that I acomplished them so well: 1. my first goal was to comapre the mean and the median sentiment of the hashtags in order to get a true sense of which aggregate function fits best this model and while analysing the outcomes I realized that these data have more outliers than we expect. Some of the median sentiments for some hashtags are less than the mean which meke sense because measuring with median gives us more accurate information about the data than mean when we have outliers in our data.

2. my second goal was to investigate which hashtags correspond to negative polarity. In my investigation I found that hashtags like debt, abortion, justice, global warming etc. asigned with negative polarity which totally make sense because of the current political conflicts around the world and in the US.

I learned a lot from this homework, however, specifically, I found that text analysis is one of the best ways to extract a lot of information that we can't get it through surveys or sampling in a natural way.