Sentiment Analysis Project

By Brian Mallari

Project Description

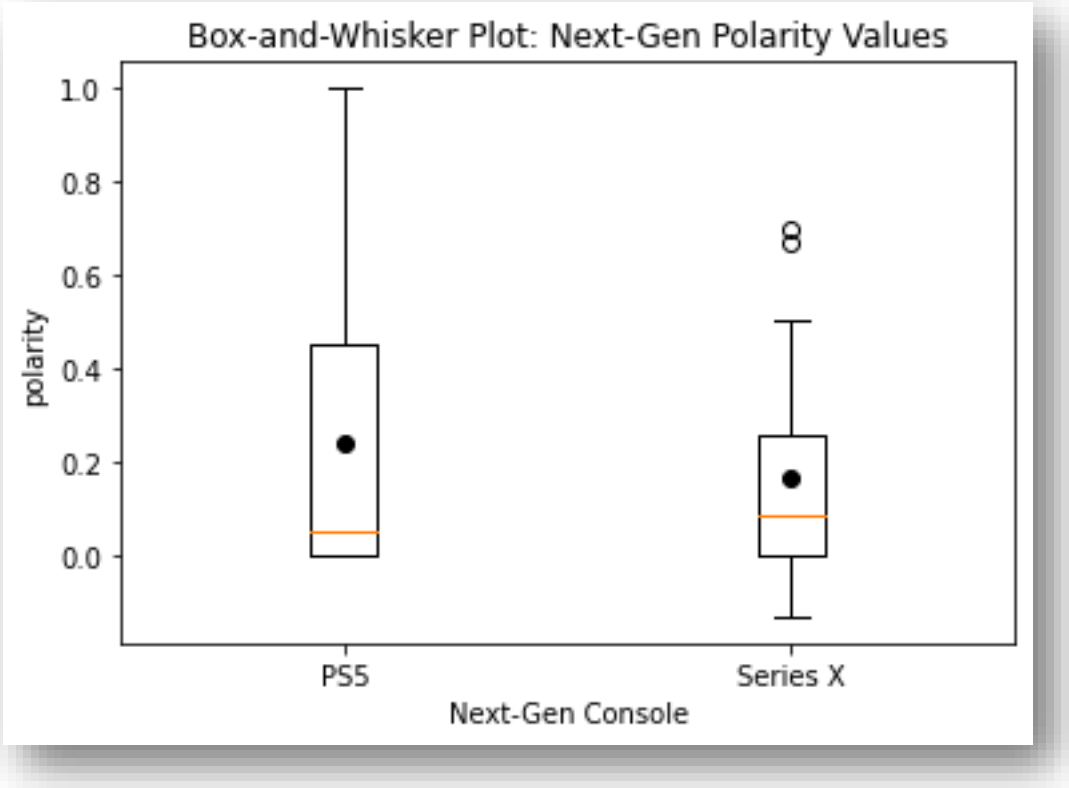
The aim of this project is to apply sentiment analysis on Tweets gathered with the Twitter API in order to answer the following question:

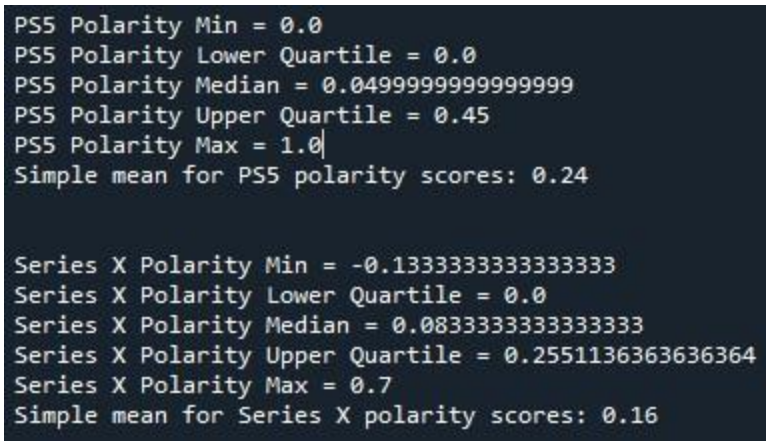
"Which next-gen gaming console should I acquire – the PlayStation 5 or the Xbox Series X?"

The Python script used for this project was based on an assignment for a class taught by Dr. Chirag Shah, a professor at Rutgers University, back when I was still a graduate student.

Data Analysis

Now since we're trying to figure out what which next-generation console to acquire based off of other people's opinions, a good place to start looking at the polarities of the Tweets for each console.



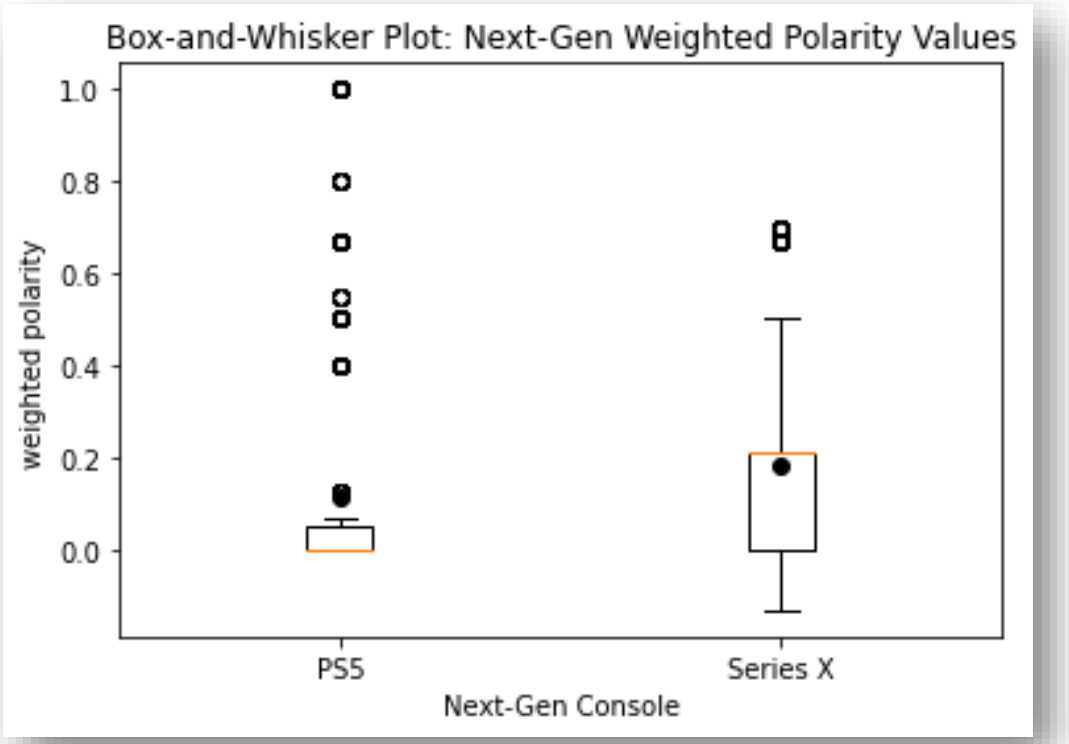


Here, we can see here that in the case of the PS5, the minimum value, the lower quartile value, and the median values are clustered together near 0. This means that half of the unique Tweets about this game system in this set of data tend to be neutral in nature. The interquartile range for the PS5 polarity values is 0.45, and the overall range is 1.0 with a maximum value of 1.0. These show that some Tweets tend to be positive in nature, which pushes the simple mean above the median.

Now for the Series X, the minimum value is actually negative which would indicate that at least one Tweet has something negative to say that relates to the system. The lower quartile value and the median value are close to 0 which would imply that at least a quarter of the unique Tweets for the Series X tend to be neutral in nature. The interquartile range for Series X polarity values is a little under 0.62 and the overall range is a little over 0.83. These values show, like in the case of the PS5, that some Tweets tend to be positive in nature, which also pushes up the simple mean of Series X polarity values above the median.

Because the two box-and-whisker plots for each set of values don't look symmetrical, especially with the presence of outliers which can skew the averages, it would be better to compare the median values. And even though at least one unique Tweet relating to the Series X has a negative polarity value, the median value for the Series X is higher than the median value of the PS5, **it would appear that based off of this data set the Series X would be the next-gen system to acquire**.

Now, worth noting here is that the two distributions of polarity values for each gaming system treat each value as though they all carry the same weight as any other value for their respective system. In other words, a high-polarity Tweet for one console can be seen as having the same degree of prevalence as a low-polarity Tweet for the same console. However, people tend to share Tweets because those people feel that the Tweet will resonate with those who will be receiving the shared Tweet, so it makes more sense to weigh each polarity score with the corresponding retweet value. That way, a highly circulated Tweet can be more telling of what people think about the gaming console in question than just the polarity value alone.




```
PS5 Polarity Weighted Min = 0.0
PS5 Polarity Weighted Lower Quartile = 0.0
PS5 Polarity Weighted Median = 0.0
PS5 Polarity Weighted Upper Quartile = 0.0499999999999999
PS5 Polarity Weighted Max = 1.0
Weighted mean for PS5 polarity scores: 0.11


Series X Polarity Weighted Min = -0.1333333333333333
Series X Polarity Weighted Lower Quartile = 0.0
Series X Polarity Weighted Median = 0.2083333333333333
Series X Polarity Weighted Upper Quartile = 0.2102272727272727
Series X Polarity Weighted Max = 0.7
Weighted mean for Series X polarity scores: 0.18
```

Now here we can see that in the case of the PS5, the weighted lower quartile, weighted median, and weighted upper quartile all are closet to 0. This means that at least 75 percent of all of the Tweets in this weighted set of polarity values tend to be neutral in nature. The overall range remains at 1.0 with the maximum still at 1.0; however, the interquartile range is now reduced to just under 0.05. The weighted mean is above the median like in the case of the simple mean in the previous observation; however, the weighted mean is 0.11 as opposed to the unweighted mean which is 0.24.

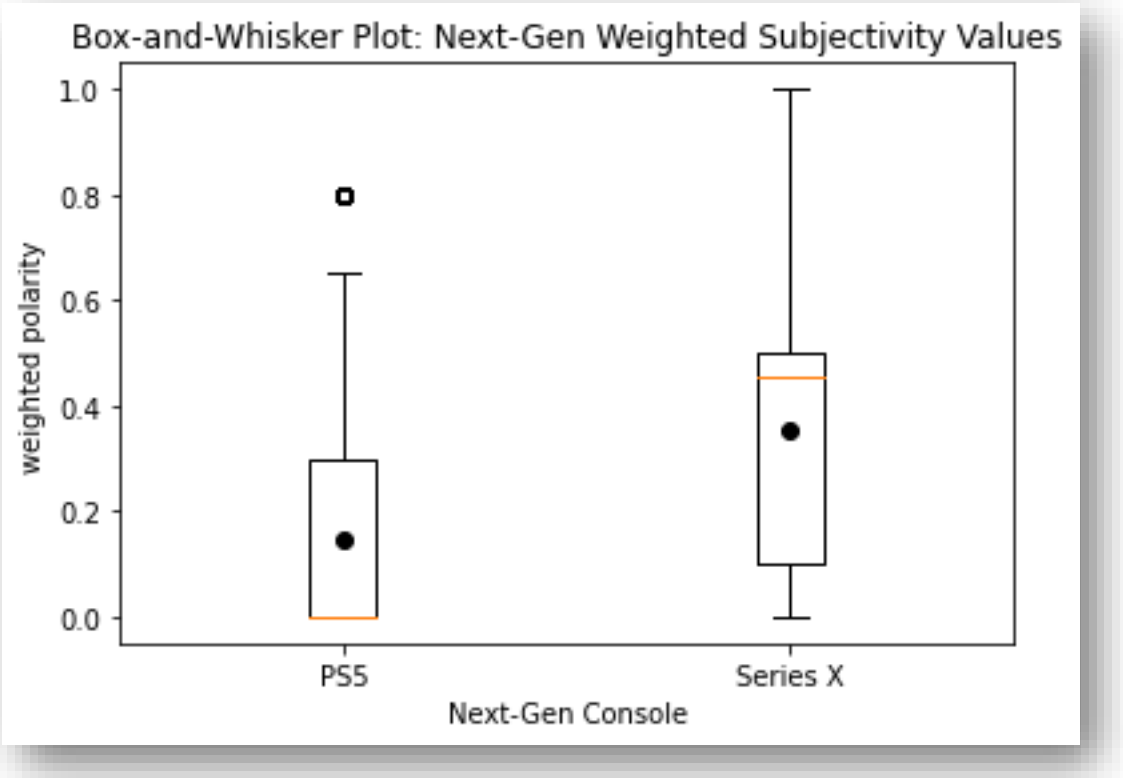
As for the Series X, the weighted lower quartile remains the same at 0.0 as the unweighted lower quartile. The weighted median is under 0.21 (as opposed to just over 0.08 for the unweighted median). Also, the weighted upper quartile is a little over 0.21 (as opposed to just over 0.25 in the unweighted upper quartile). The overall range remains at a little over 0.83, and the interquartile range for this set of data is just over 0.21. However, the greater weighted median value over the unweighted median value would indicate that the central tendency of the data shifts upward, which is reflected in the weighted mean being 0.18 as opposed to the unweighted mean which is 0.16.

Now just like before, the two box-and-whisker plots for the sets of unweighted values don't look symmetrical, especially with the presence of outliers which can skew the averages. Therefore, it would be better to compare the median values once again. The weighted median value for the Series X is higher than the weighted median value of the PS5, **so it would appear that based off of this data set the Series X would still be the next-gen system to acquire**.

Some Considerations

Subjectivity Scores

Subjectivity, which is a measure of how objective or subjective a Tweet is, was also calculated based off of the content of the Tweets.



```
PS5 Subjectivity Weighted Min = 0.0
PS5 Subjectivity Weighted Lower Quartile = 0.0
PS5 Subjectivity Weighted Median = 0.0
PS5 Subjectivity Weighted Upper Quartile = 0.3
PS5 Subjectivity Weighted Max = 0.8
Weighted mean for PS5 subjectivity scores: 0.14


Series X Subjectivity Weighted Min = 0.0
Series X Subjectivity Weighted Lower Quartile = 0.1
Series X Subjectivity Weighted Median = 0.4545454545454545
Series X Subjectivity Weighted Upper Quartile = 0.5
Series X Subjectivity Weighted Max = 1.0
Weighted mean for Series X subjectivity scores: 0.36
```

People tend to share Tweets because those people feel that the Tweet will resonate with those who will be receiving the shared Tweet. Therefore, like in the case of the weighted polarity values, it makes sense to weigh each subjectivity score with the corresponding retweet value. That way, a highly circulated Tweet can be more telling of whether what people think of the gaming console in question is more objective or more subjective in nature.

Now in the case of the PS5, the weighted minimum, the weighted lower quartile, and the weighted median are all set to 0.0; this means that at least half of all of the Tweets in this data set tends to have objective content. The interquartile range is 0.3, and while the total range of values for this set is 0.8, **overall, the Tweets for the PS5 are objective in nature**, especially with a weighted subjectivity score of 0.14.

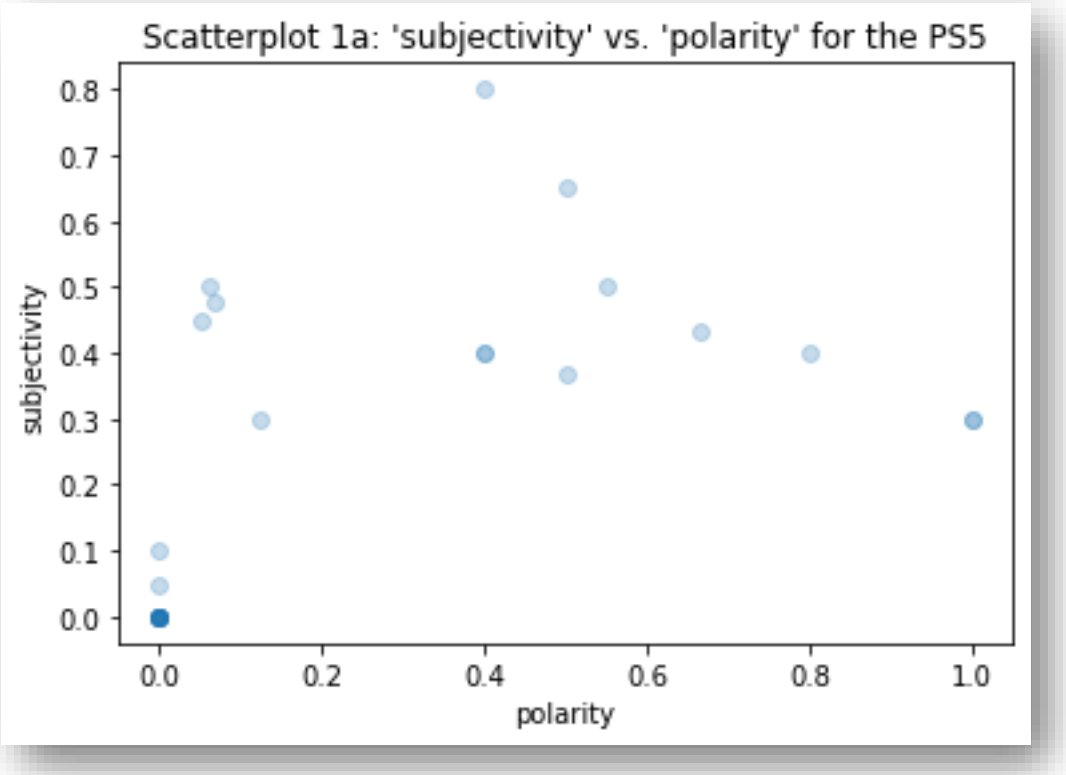
And in the case of the Series X, the weighted minimum is 0.0. However, the weighted lower quartile is 0.1, and the weighted median is just over 0.45. The interquartile range is 0.4, and the weighted maximum is 1.0. The interquartile range is 0.4, and the total range of values for this set is 1.0, **so overall the Tweets for the Series X are more subjective than the Tweets for the PS5. However, with a weighted mean of 0.36, the Tweets for the Series X could still be considered to be tending towards objective in content**.

Polarity vs. Subjectivity Scores

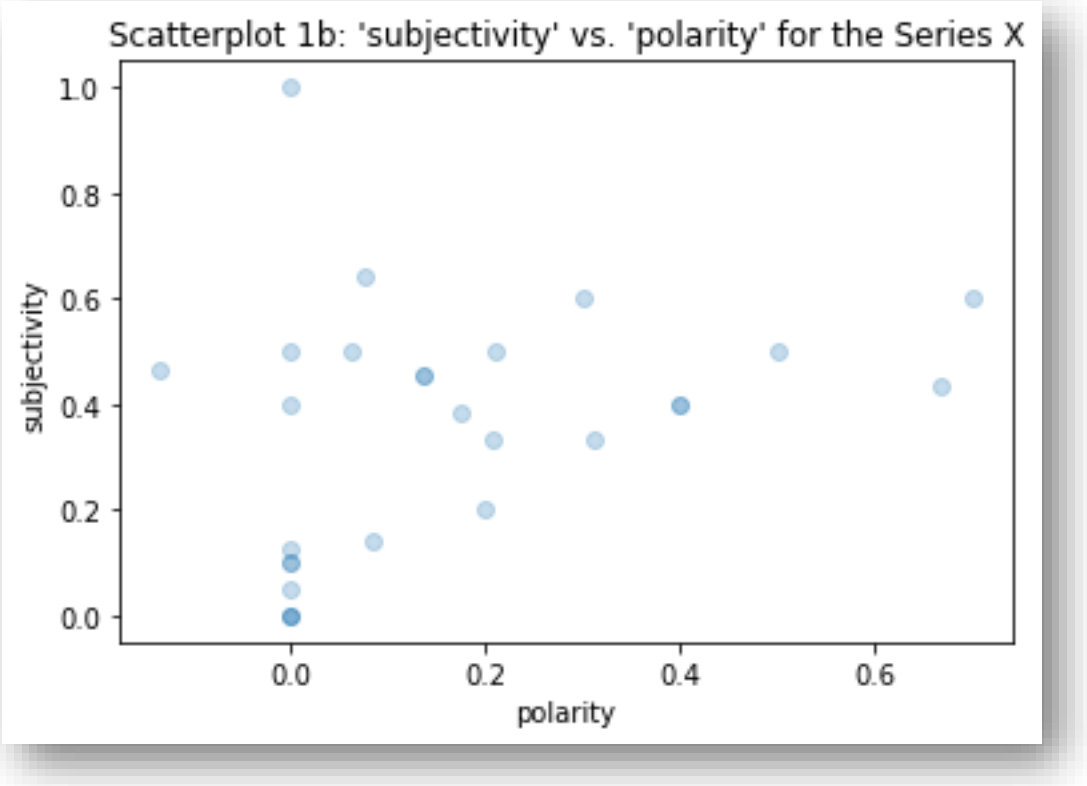
```
Correlation matrix for variables associated with Tweets relating to the PlayStation 5
              author id     retwc  followers   friends  polarity  subjectivity
author id      1.000000 -0.129843  -0.109122 -0.044112 -0.070763      0.050750
retwc         -0.129843  1.000000  -0.139239  0.138907 -0.312993     -0.312439
followers     -0.109122 -0.139239   1.000000 -0.047605 -0.267615      0.015967
friends       -0.044112  0.138907  -0.047605  1.000000  0.251763      0.157099
polarity      -0.070763 -0.312993  -0.267615  0.251763  1.000000      0.576255
subjectivity   0.050750 -0.312439   0.015967  0.157099  0.576255      1.000000


Correlation matrix for variables associated with Tweets relating to the Xbox Series X
              author id     retwc  followers   friends  polarity  subjectivity
author id      1.000000  0.078611  -0.244841 -0.112226 -0.255441     -0.336594
retwc          0.078611  1.000000  -0.071812 -0.024234  0.041718     -0.002818
followers     -0.244841 -0.071812   1.000000 -0.049853  0.046160      0.308975
friends       -0.112226 -0.024234  -0.049853  1.000000  0.278072      0.110425
polarity      -0.255441  0.041718   0.046160  0.278072  1.000000      0.315375
subjectivity  -0.336594 -0.002818   0.308975  0.110425  0.315375      1.000000
```

Based off of the correlation table for the unweighted values, both consoles have positive, linear relationships between the polarity and subjectivity values (0.576225 and 0.315375 for the PS5 and Series X respectively).



For this scatterplot, the dots were made slightly transparent so that any overlap can be depicted by darker dots. Now in the case of the PS5, there is a dark dot towards the lower, left-hand side of the scatterplot which would imply that the content of several Tweets would be declarative in nature (i.e., the Tweet is just a statement of some fact or detail). This concentration of dots towards the bottom corner of the scatterplot could account for the correlation value of over 0.57 between the unweighted polarity value and unweighted subjectivity value. The other dots outside of the lower left-hand corner appear to be spread out.



Like in the scatterplot for the PS5, the dots were made slightly transparent so that any overlap can be depicted by darker dots. Now in the case of the Series X, there appears to be a concentration of dots towards the bottom of the subjectivity scale, though not necessarily to the same degree as the PS5, since the dot at the very bottom of the Series X boxplot isn't as dark as the bottom-most dot in the PS5 scatterplot. This indicates less overlap of dots for the Series X in that region of the scatterplot. Instead, the dots look more scattered about in the case of the Series X, which would account for the correlation value of over 0.31 between the unweighted positivity value and the unweighted subjectivity value. The larger spread of dots in the case of the Series X would be indicative of excitement (i.e., positive opinion) with respect to this gaming system.

Now unlike the case of just polarity values and just subjectivity scores where the values were weighed by the retweet count, the correlation coefficient between these two fields or categories was based off of the unweighted value. That is, the polarity score of one Tweet was correlated to its subjectivity score, the polarity score of a second Tweet was correlated to its subjectivity score, and so on and so forth. This is because both the polarity score and the subjectivity score are based off of the content of Tweet, and

retweeting a Tweet shouldn't change the content of the original Tweet, even if the sentiment of the Tweet is shared by all of people who've retweeted it.

Data with Respect to Time

The data were collected back in January of 2021, so it is possible that the polarity values would change over time. Therefore, it might be worthwhile to collect these sentiment values over time and then to plot them in a line graph to identify any trends.

It might also be worthwhile to track the number of units sold for each next-gen console over time and then run a correlation analysis with the sentiment values. If there is a strong, positive correlation between sentiment values and sales values, then the sales value might serve as social proof that one gaming console is more preferable than the other if people are willing and able to spend their money on the unit. However, the number of units sold on a daily basis might not be easily available to the public, so this sort of analysis might not be viable in the case of next-gen gaming consoles – unless, of course, if this analysis is done in the context of employment with Sony or Microsoft where daily sales data might be more easily accessible, or if this analysis is performed on a good or service for which sales data are more easily accessible.

Factors for Purchasing Gaming Consoles

While the weighted median polarity value for the Series X is higher than the weighted median polarity value for the PS5, polarity values alone might not be the sole driving force for purchasing one console over another. For example, some games may end up being exclusive to just one of the consoles, such as Final Fantasy VII Remake which will be only on the PlayStation, so if someone wants to play that game, then they'll more likely to purchase the PlayStation over the X-Box despite the results of any sentiment analysis. A person's decision to purchase one console over another might be dictated by what people in their own social group will purchase since using the platform could equate to a shared experience for all of the members in the group, even if a game could be cross-platform (i.e., the game can be played simultaneously online using different platforms).

A person's decision to acquire one console over another could dictated by past purchases. For example, if someone has acquired prior generations of the PlayStation, such as the PlayStation 2 or the PlayStation 3, then that person might be inclined to purchase the PlayStation 5 because they're familiar with the Sony brand and the subsequent gaming experience. And then there are some people out there who are willing and able to purchase both next-gen gaming console, such online content creators who focus on video game gameplay or commentary and are open to playing on a wide variety of platforms.