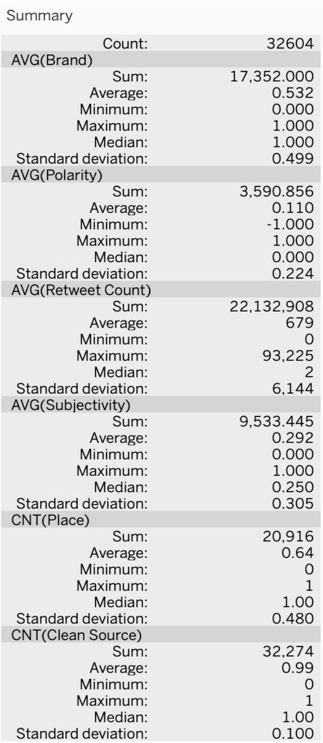
Tiffany Elamparo, Kelly Faenza
USD School of Business
Professor Sanchez
ECON 385: Business Analytic Strategy
15 September 2020

A Comparison of Sentiment Toward Apple and Samsung Using Twitter Data

I. Descriptive Statistics Table



Summary

Count:	32604
AVG(Brand)	
Sum:	17,352.000
Average:	0.532
Minimum:	0.000
Maximum:	1.000
Median:	1.000
Standard deviation:	0.499
AVG(Polarity)	
Sum:	3,590.856
Average:	0.110
Minimum:	-1.000
Maximum:	1.000
Median:	0.000
Standard deviation:	0.224
AVG(Retweet Count)	
Sum:	22,132,908
Average:	679
Minimum:	0
Maximum:	93,225
Median:	2
Standard deviation:	6,144
AVG(Subjectivity)	
Sum:	9,533.445
Average:	0.292
Minimum:	0.000
Maximum:	1.000
Median:	0.250
Standard deviation:	0.305
CNT(Place)	
Sum:	20,916
Average:	0.64
Minimum:	0
Maximum:	1
Median:	1.00
Standard deviation:	0.480
CNT(Clean Source)	
Sum:	32,274
Average:	0.99
Minimum:	0
Maximum:	1
Median:	1.00
Standard deviation:	0.100

For our sentiment analysis, we are comparing two phone brands: Apple and Samsung. The variables of interest are brand, polarity, retweet count, subjectivity, place, and clean source. Brand is a dummy variable that indicates observations as either 0 for tweets containing Apple keywords and 1 for tweets containing Samsung keywords.

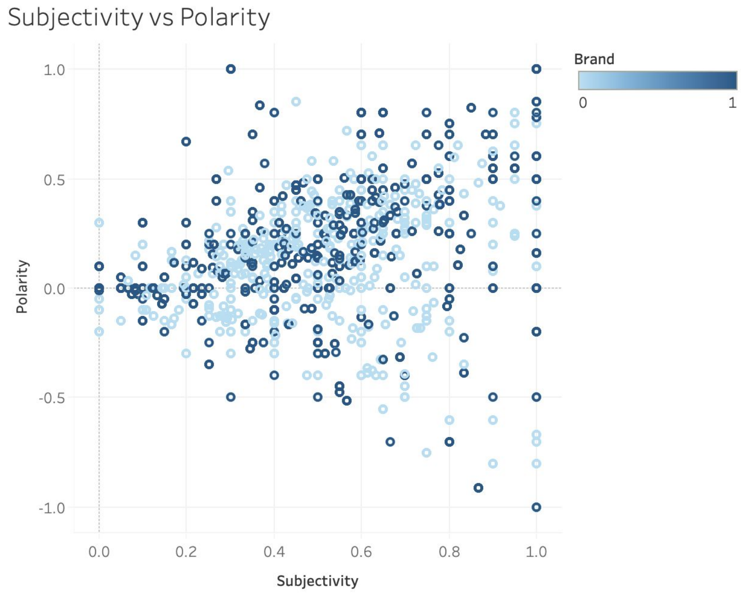
II. Sampling Method and Frame

We chose Apple and Samsung because they are two leading tech companies and most people use one of these brands in their daily life. Many people have strong opinions about which products are more user friendly or financially accessible, and we wanted to see if these sentiments would show up in Tweets. When gathering our sample, we looked for tweets containing the company names, the names of their best selling products, and the names of their CEOs. We chose these keywords because we wanted to capture sentiment toward the brand as a whole, products, and the leadership of the company. We cleaned the data in Python, which involved tasks such as removing emojis and limiting the language to English. The clean code is the data that we used as our sample.

III. Comparison and Explanation

Brand Legend:
0= Apple (light blue)
1= Samsung (dark blue)

1: Relationship between Subjectivity and Polarity

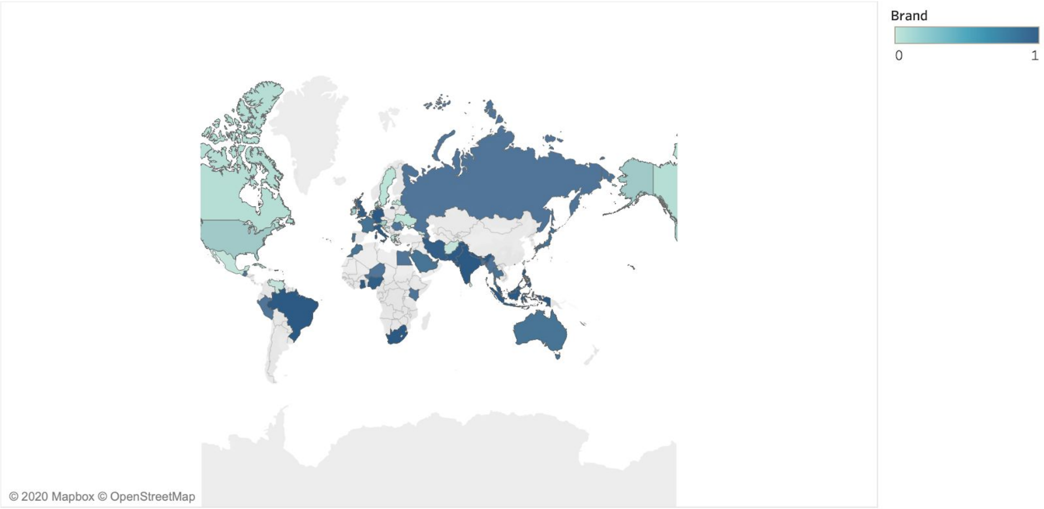


Subjectivity vs. Polarity. Color shows Brand.

In this graph, we compare the polarity and subjectivity of tweets regarding Apple and Samsung. We noticed that the tweets with low subjectivity also have low polarity, which is what we would expect since subjectivity is a measure of being based on feelings or opinions rather than on facts. The tweets that are based on facts are more neutral in terms of polarity than the highly subjective tweets. As subjectivity of the tweets increases, the range of polarity also expands because the polarity of the tweets mentioning these brands are both negative and positive. It also appears that there are more observations in the range of positive polarity than negative polarity, meaning on this topic, people generally tweet more positive than negative things when discussing these brands. Samsung is mentioned by twitter users in the most polarizing tweets. We observed that the dark blue color, which represents Samsung, is the subject of the tweets closest to 1 and to -1. It does not appear that either firm is more favored than the other because Samsung is the subject of both the most negative and positive tweets.

2: Map of Subject of Tweets by Location

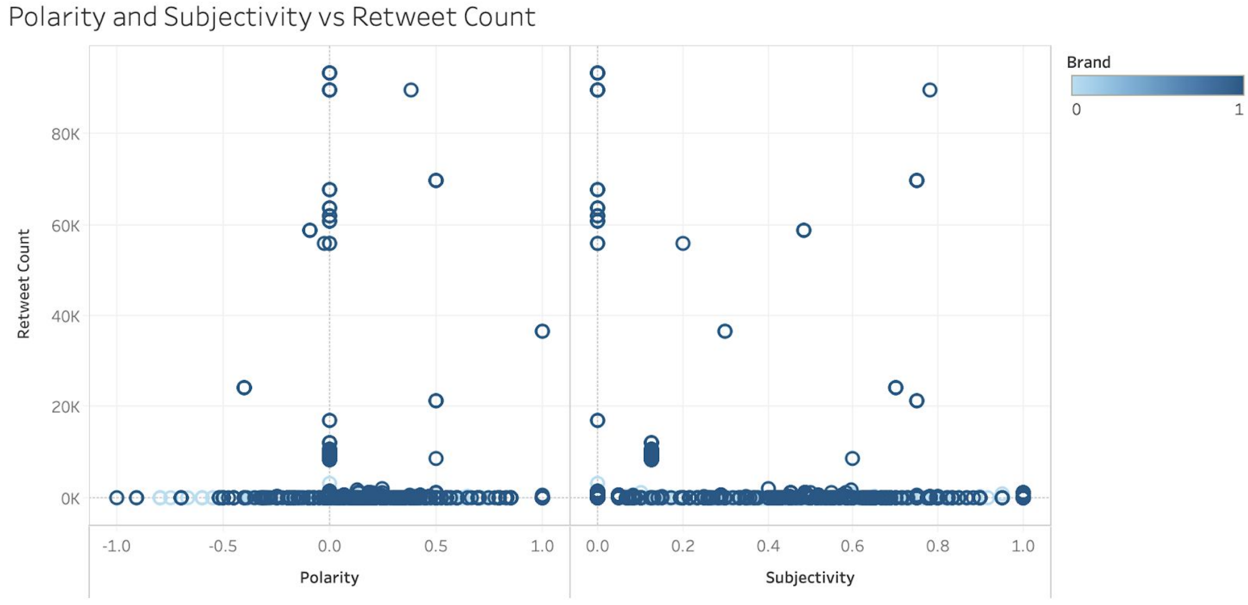
Map



Map based on Longitude (generated) and Latitude (generated). Color shows Brand. Details are shown for Place.

This map depicts the geographic location of the extracted tweets by brand. The color seen in North America is a moderate shade of blue that is neither fully light or dark, showing that both brands are present on that continent. However, the color is arguably more on the lighter side, which means that Apple has a slightly larger presence than Samsung. On the other hand, Samsung is significant in countries such as Russia, Brazil, and India. Australia also appears to have a majority of Samsung Twitter users, but we can see that Apple users are still present given how muted its shade of blue is in comparison. We would predict that Apple and Samsung products sell best in the areas on the map that they are mentioned in, but we could further this prediction by comparing the map with the sentiment analysis of each Tweet. From the map, it is unknown if either is more favored than the other, as both brands appear to have an overall equal presence on a global scale in countries where location of tweets are listed. In areas that have neither Samsung or Apple present, it is possible that these brands are prohibited from the market, or even that Twitter is banned from the country. For example, in China, the cell phone market is largely controlled by Huawei so Apple and Samsung are absent. Even if Apple and Samsung were permitted to enter the market, the Twitter application is banned from use.

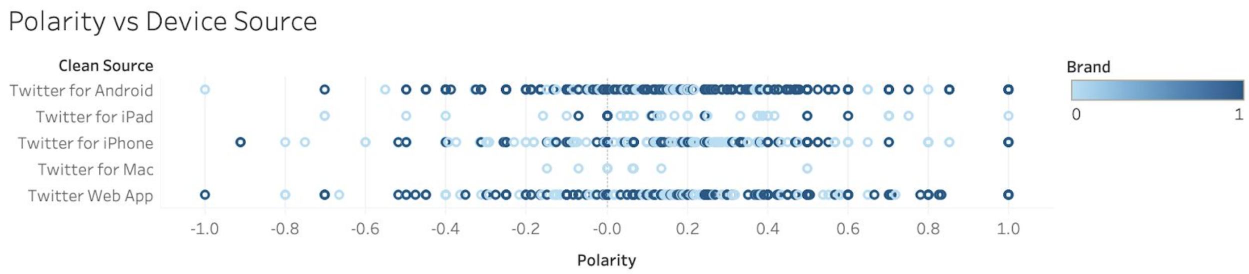
3: Relationship between Polarity and Subjectivity with Retweet Count



Polarity and Subjectivity vs. Retweet Count. Color shows Brand.

In our graph examining the relationship between retweet count with polarity and subjectivity, we found that the highest amount of retweeted tweets are the ones that are least polarizing and least subjective. People share the tweets that are factual rather than opinion based and tweets that are more neutral in the message they say about the brand. In other words, the highest amount of retweets are seen in observations with subjectivity and polarity scores close to zero. It also seems that the majority of retweets contain Samsung, as all of the light blue observations, the Apple observations, appear to be under the one thousand retweet mark for both subjectivity and polarity. Twitter users are more engaged with Samsung because tweets regarding this brand are more likely to be retweeted. Although the majority of observations for polarity appear to be clustered around the zero mark, there are arguably more positive than negative tweets. It can be inferred that users are more inclined to retweet neutral tweets over positive tweets, and positive tweets over negative tweets. An interesting phenomenon is seen on the zero line for both polarity and subjectivity, as there seems to be overlapping coordinates. For example, there is an observation that is exactly (0, 19K) seen on polarity and subjectivity. These tweets may just be facts regarding the company because many have a neutral polarity, but it is clear that information regarding Samsung is being distributed more on Twitter.

4. Polarity of Tweets Based on Device Source



Polarity for each Clean Source. Color shows Brand. The view is filtered on Clean Source, which keeps Twitter for Android, Twitter for iPad, Twitter for iPhone, Twitter for Mac and Twitter Web App.

This is a graph of the relationship between the polarity of a tweet and the device that sent the tweet. We found this interesting because we wanted to know if it was users of the brand that have strong feelings about it or users of the opposing brand. The Source being Twitter for iPad, Mac, and iPhone implies that the twitter user is an Apple user. It appears that most of the tweets made from these devices are about Apple. The polarity of the tweets seem to be balanced between positive and negative values, yet for the source Twitter for iPhone, the most polarizing tweets are both about Samsung. The source Twitter for Android is made up of all users of phones with Android operating systems, yet this includes Samsung users. Both brands are the subjects of tweets from this device and the most positive tweet is about Samsung, while the most negative is about Apple. We had to clean this data to remove the many bots that existed to retweet messages containing both Apple and Samsung. We found bots such as “Apple Retweet Bot” that were not representative of actual people, so we removed this data from this graph. On Twitter for Android, Samsung is favored, on the Apple devices overall Apple is favored, and on the Twitter Web App, which is used by everyone not just users of a certain brand, Samsung is more favored but also is the subject of the most negative tweets.

IV. Limitations

The tweets gathered only included English speakers because we specified the code to do so. Another limitation is that the location service is disabled in some countries, so when we examine the map, it only includes the locations of tweets where geographic location is enabled. For example, we see this limitation in South Korea because Samsung is a Korean company, yet there are no tweets about either company on the map.

Other limitations of our sentiment analysis is that we are only sampling from twitter users over a 3 day range. This is a problem because twitter users are not representative of the world and we cannot base our understanding of public sentiment based on the popular sentiment on twitter. When we made a prediction for figure two about where each company would perform best, it should be acknowledged that representation bias may exist. The small date range is also a problem because we are unable to determine sentiment before and after events that would get people discussing these brands, such as a new product being released. If the time period were longer, we would gain a better understanding of twitter users sentiment toward the brands because this could be dependent on the time of year or current events.

We also faced the problem of gathering data from dummy bots. These are not representative of actual people and therefore interfere with our sentimental analysis. These bots are created to retweet data or create data, yet their tweets are still recognized as being subjective and polarizing. Since these are bots, they are not actually expressing sentiment, so their inclusion in our dataset is a limitation of our analysis because they are outliers in subjectivity and polarity. We excluded these when examining polarity based on device source, but they were still included in the rest of the analysis.

Although we observed relationships between our variables, these variables are correlations and we cannot prove causation. We are recognizing and analyzing patterns, but cannot confirm what causes the variables to behave this way.

If we were to continue this analysis, we would include more keywords because we only included four key words for each brand. We would be able to gather more insight if we included more keywords, so the small number used is a limitation of our analysis.

These are just a few of the limitations that we faced, yet despite these limitations, we were able to extract sentiment from tweets and gain an understanding of the relationships between sentiment, polarity, location, retweet count, and source of device used.