95-851: Data Science for Product Managers
NLP Analysis of Consumer posts on Samsung Galaxy S8 vs. iPhone 8, iPhone X

Nidhi Bhaskar nidhibha@andrew.cmu.edu
Woojin Park woojinpa@andrew.cmu.edu
Aditya Soni adityavs@andrew.cmu.edu
Vishal Ramesh vishalra@andrew.cmu.edu

Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, USA

ABSTRACT

This report will primarily focus on the analysis that we performed on social media posts to determine the sentiments of people towards two primary smartphone manufacturers, namely Samsung and Apple. To be more specific, we will be obtaining sentiment information on the Samsung Galaxy S8, which was released on March 29th, 2017, and the iPhone 8 and X, which were released together on September 22 2017. We will be looking to see the opinions of people across different countries through their comment/posts/reviews on various social media networking platforms pertaining to Galaxy S8 before its releases, and if it was affected in any way post its release, and post the release of the 2 iPhone versions respectively. We will also be looking at the sentiments towards the iPhone versions before and after its release separately as well.

The data is obtained from multiple sources, some of which are Twitter, Tumblr, and Reddit. The data is spread through March till November, with portions missing in the middle (post the release of the Galaxy S8 and before the release of the iPhones).

INTRODUCTION

We evaluate the performance of each of the smartphone models, i.e., Samsung Galaxy S8, iPhone8 and iPhone X by analysing how people feel about each of the models before and after the release dates. We do this by gathering data from online sources across various social networking platforms in the form of posts/comments/reviews. We majorly make use of Sentiment Analysis, Topic Modelling (NMF), Time Series Prediction Modelling as the different techniques to understand what people feel about each of the products thereby leading us to a good model that validates the findings.

We find the most important features that people focus on while they comment or post their opinions about the product and see if there is a positive, negative or neutral sentiment attached to it. By doing this, we can find out how the companies can improve their businesses by focusing on features that people feel the product has missed out on in their upcoming versions/releases.

We also look at the most important attributes such as “Price”, “Quality” and “Value” to see how people feel the products has performed individually. We draw a comparison between the smartphone models to observe which product is valued the highest, is priced appropriately and has the best quality.

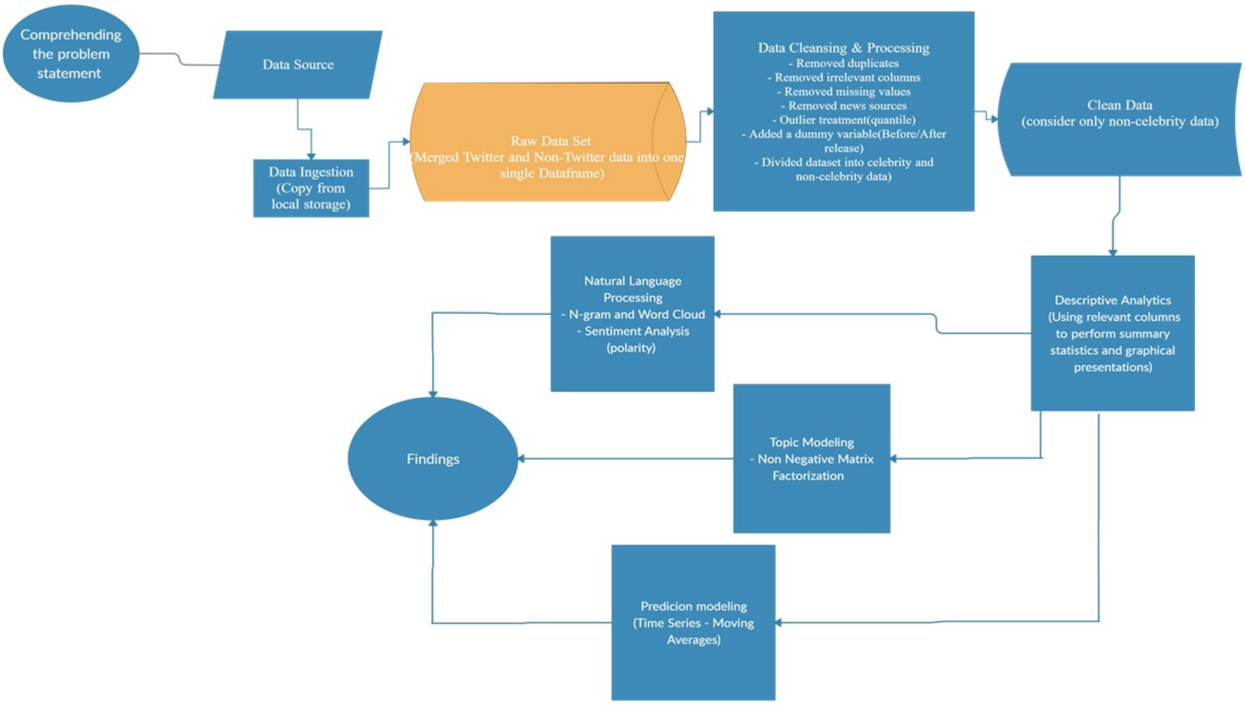
By plotting the trends of sentiments in terms of positivity, negativity and neutrality using polarity for individual products before and after the release, we predict the uptake/adoption of the product.

SCOPE OF THE DATA

The data that we are provided with comes from different social networking platforms which consists of comments/reviews/posts by people spread across the world belonging to different professions, age ranges, gender, etc. The complete dataset consisted of 500,000+ in separate files that required to be merged. At the first glance of the dataset, we can observe a lot of irrelevant columns, missing values and duplicate records that had to be taken care of. We also see that there a lot of records from news sources/celebrities who may have potential collaborations with the smartphone companies, and this might result in our analysis being highly biased towards a single sentiment. Once we have taken care of the unclean dataset and have the most important and processed data, we can move on to the analysis phase.

DATA ANALYSIS PIPELINE

As an analysis pipeline (a) we formulate our problem statement, (b) and collect and merge the data (c) perform data cleaning & pre-processing, (d) descriptive analysis (e) perform analysis (Natural Language Processing, Topic Modeling, Prediction- Time Series) and (f) report findings.


Comprehending the problem statement
Data Source
Data Ingestion (Copy from local storage)
Raw Data Set (Merged Twitter and Non-Twitter data into one single Dataframe)
Data Cleansing & Processing
- Removed duplicates
- Removed irrelevant columns
- Removed missing values
- Removed news sources
- Outlier treatment(quantile)
- Added a dummy variable(Before/After release)
- Divided dataset into celebrity and non-celebrity data)
Clean Data (consider only non-celebrity data)
Descriptive Analytics (Using relevant columns to perform summary statistics and graphical presentations)
Natural Language Processing
- N-gram and Word Cloud
- Sentiment Analysis (polarity)
Topic Modeling
- Non Negative Matrix Factorization
Predicion modeling (Time Series - Moving Averages)
Findings

DATA PRE-PROCESSING AND CLEANING

The dataset provided to us is split into 11 excel documents which consists of 2 twitter and rest non-twitter files. We were required to merge all of this in the same data frame and then moved on to the following steps:

- The column with the relevant text data is titled 'Sound Bite Text', which we will use to perform Natural Language Processing(NLP) techniques to get the desired sentiments about various products and how they change over time based on the release of other products.
- The data also contains replies, reposts, and reblogs. Since this might echo the initial sentiment that was present in the original post, we will be ignoring such data, and look at only the original posts. This is done by filtering based on the column titled ‘Post Type’, and the value we are looking for in this column is ‘Original’.
- It also makes sense to ignore the texts from popular social media users, and news sources, as it might have an inherent bias in it (Sponsored posts, paid reviews, etc). Thus, we will be eliminating all text that comes from users with a high number of followers and from those users whose daily page views are really high. The relevant columns to filter these posts are titled ‘No. of Followers/Daily Unique Visitors’ and ‘Author Name’. The ‘Author Name’ column is relevant as we can use that column to filter our posts from sources who have News in their title (for example ‘247 NEWS WEB’).
- There are also a lot of extra data present in the files, data like ‘Author Gender’ and ‘Author Country’. We have decided to only use certain columns for our analysis, and they are 'Post ID' , 'Sound Bite Text' , 'Source Type' , 'Post Type', 'Published Date (GMT-04:00) New York' , 'No. of Followers/Daily Unique Visitors', 'Professions' , 'Author Gender' , 'Author Name'.
- The data also contains texts that have been deleted/removed. This can be observed by the ‘Sound Bite Text’ column containing the value ‘Post deleted by the author’. These add no value to the analysis, and hence they will be removed as well.

Data Pre-processing step

- Due to the nature of tweet parsing, it is possible that there are duplicate entries. These need to be deleted in the pre-processing step as it might cause some skews later on. Our code suggests that there are 1443 duplicate entries which have all the values as common, and they were all dropped.
- In our columns of interest, if there are any NULL values, then these columns are being dropped as well.
- After this, we only have to retain the entries which are Original Posts. Filtering based on ‘Post Type’ gives us the desired results.
- Next, we will be looking at the number of followers a user has, and if it is really extreme (an outlier), their posts will be removed from our analysis. As mentioned earlier, this goes back to the concept of paid posts and bias in the text. After the outliers are removed, we group people based on the number of followers they have, and all popular users will

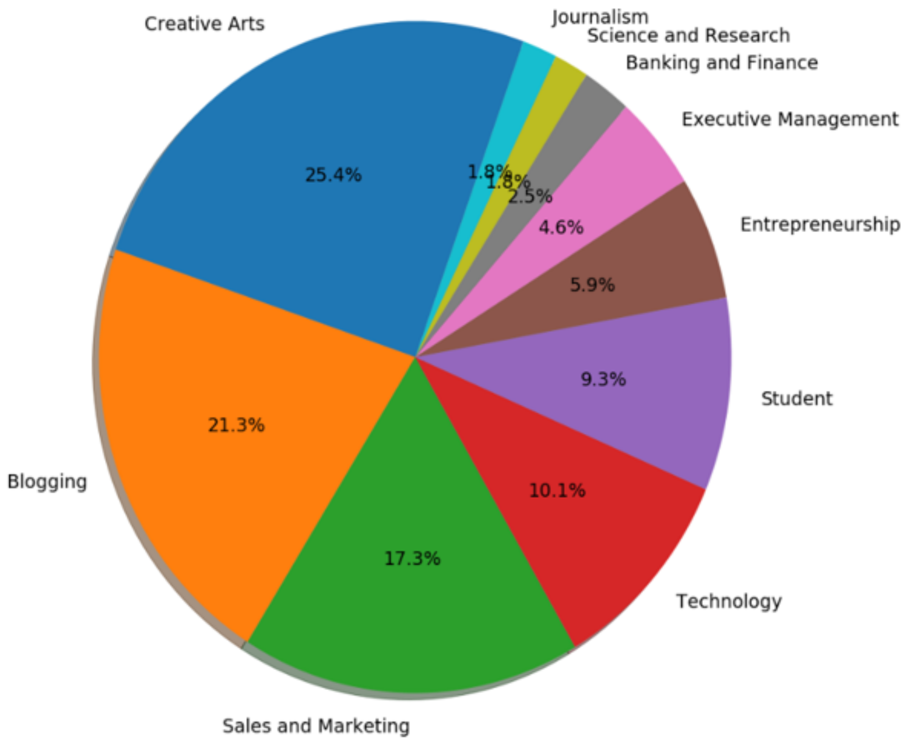
be removed. From our analysis of the user data, it shows that more than 88% of the users have 0 followers. But the users in the 99th percentile have more than 551195 followers. The users in the 95th percentile have almost 10000 followers. Thus we treat users above the 95th percentile as outliers, and we will remove their tweets. Moreover, the 90th percentile user has 226 followers while the 89th percentile has 96 followers. Thus, people who have 96 followers and above are considered popular users and their tweets will not be used for analysis in the initial stages.

  - Next, we split the data into 3 buckets
    - Data exclusively about iPhone 8
    - Data exclusively about iPhone X
    - Data exclusive about Samsung Galaxy.
  - This was necessary as we want to find the sentiment of each of these products as a stand-alone entity, to figure out the sentiments across the entire timeline we have the data for.
- Next, we split the data to pre and post release dates for the respective models.
- After all this is done, we finally have the data that we need to perform data analysis on. The resultant dataset only contains the non-celebrity records which safeguards us from any bias in our output analysis with relevant columns.

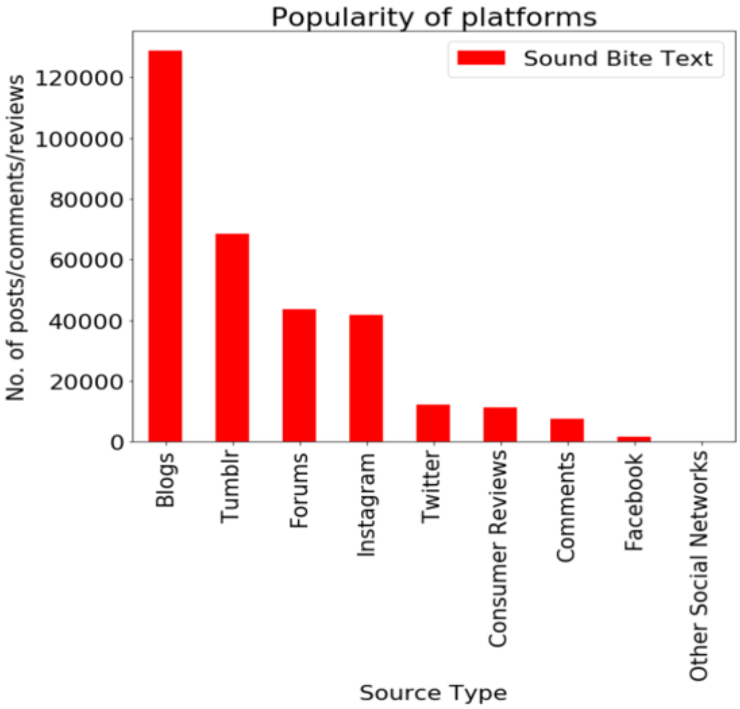
DESCRIPTIVE ANALYSIS

- The top 10 professions to which people belong to having the maximum number of comments/reviews/posts across the twitter and non-twitter platforms.

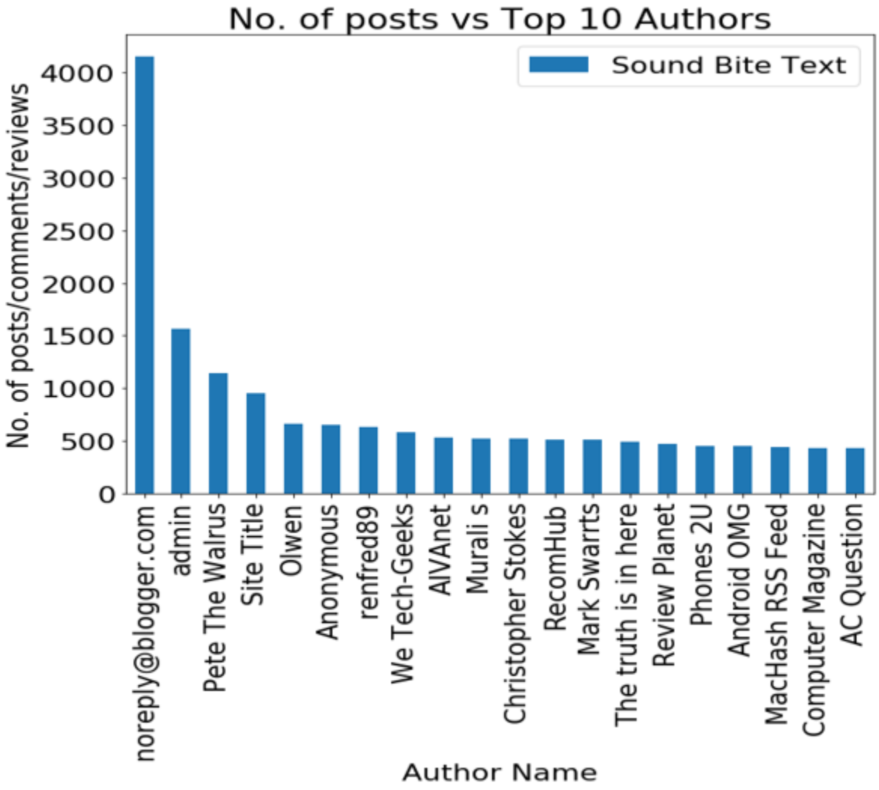
Percentage of Posts/comments/reviews by top 10 most popular Professions



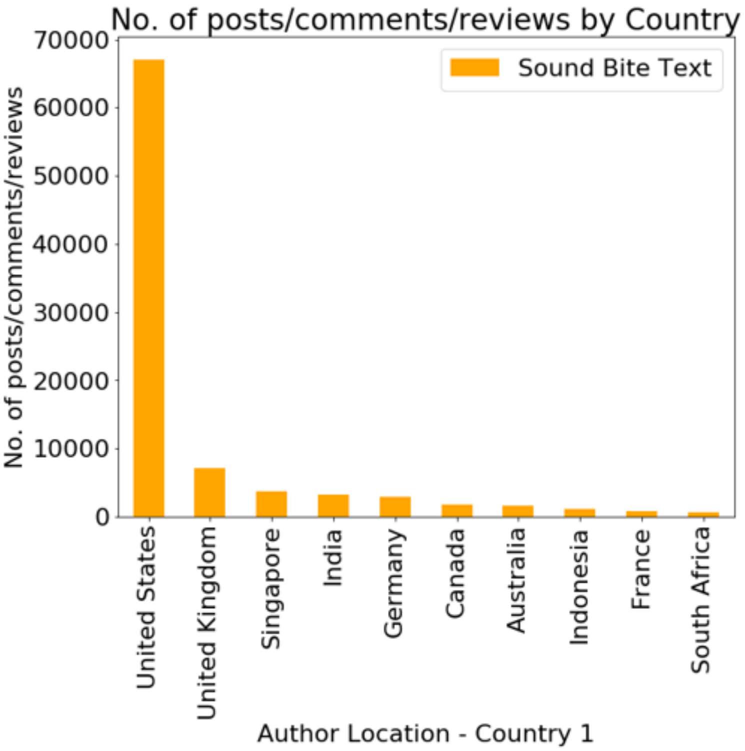
- Popularity of the platforms based on the number of posts/comments/reviews.



- Finding the top 10 authors that have the highest number of comments/posts/reviews to see that our analysis would not be biased by a single person's opinion.



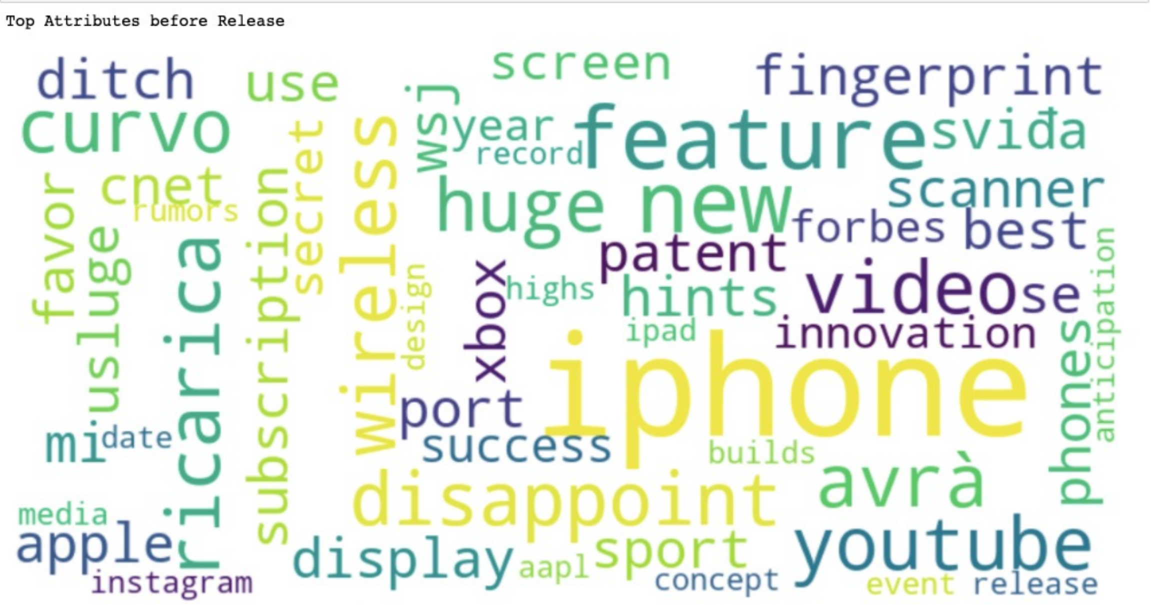
- Displaying the top 10 countries from where we get out the maximum comments/reviews/posts from people.



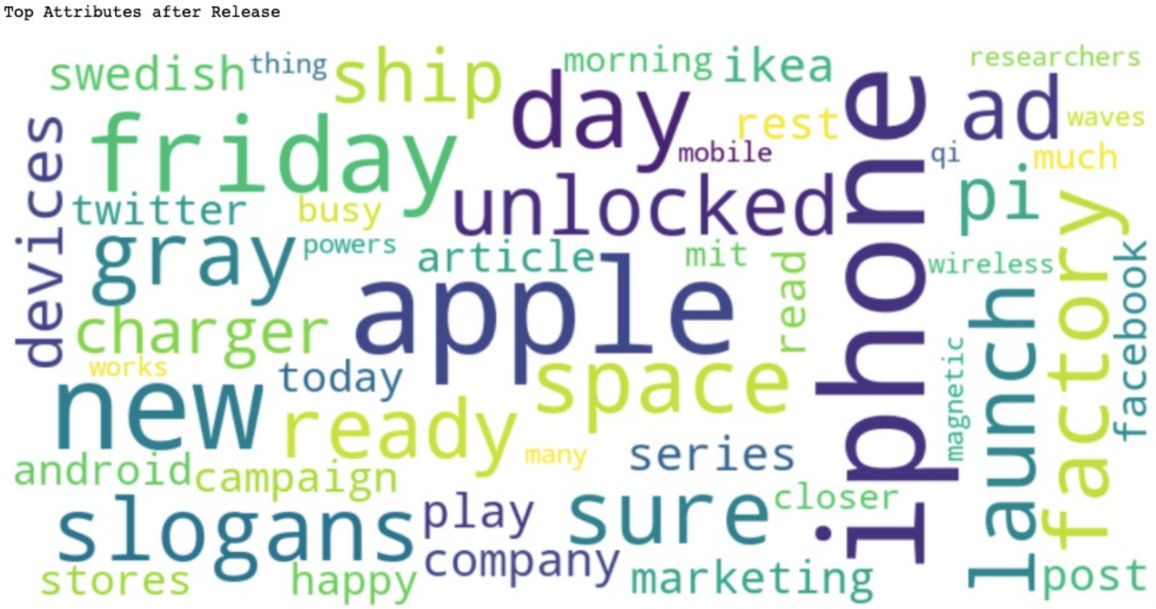
EXTRACTING THE MOST IMPORTANT ATTRIBUTES

First, it is important to understand what people have been talking about the respective products before and after release. This is just to get an understanding of the most common talked about attributes. It must be noted that in this step, we are not focusing on the sentiment of what each tweet/blog represents, but rather just what people are talking about (i.e. what the most frequent words are). Our analysis revealed the following data, which we have represented using word clouds.

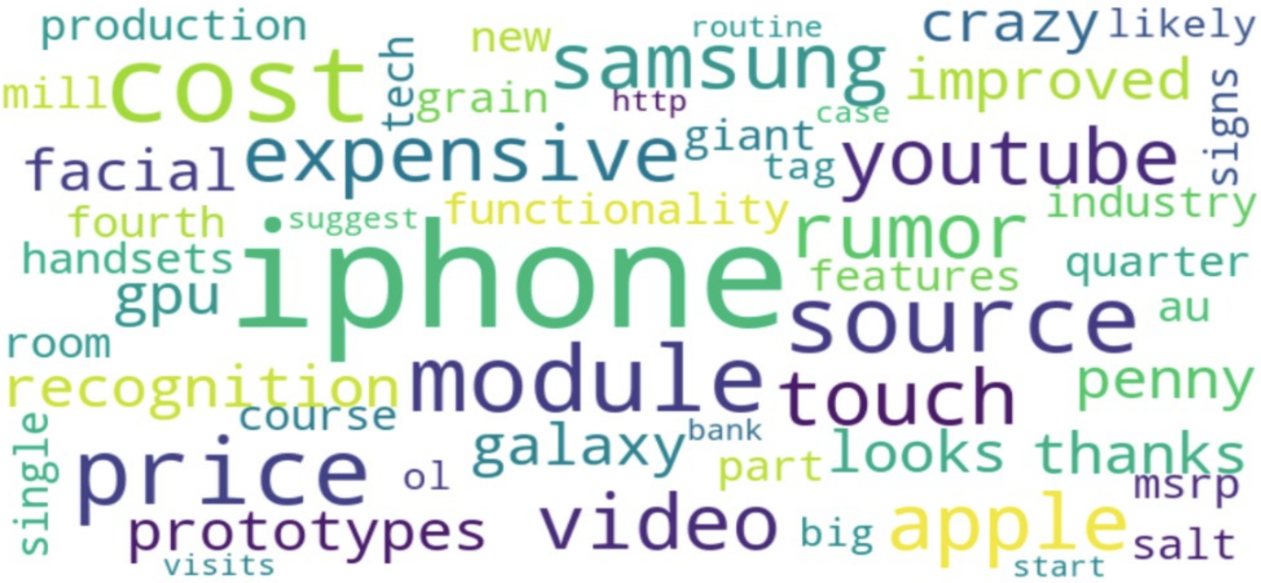
iPhone 8 before release



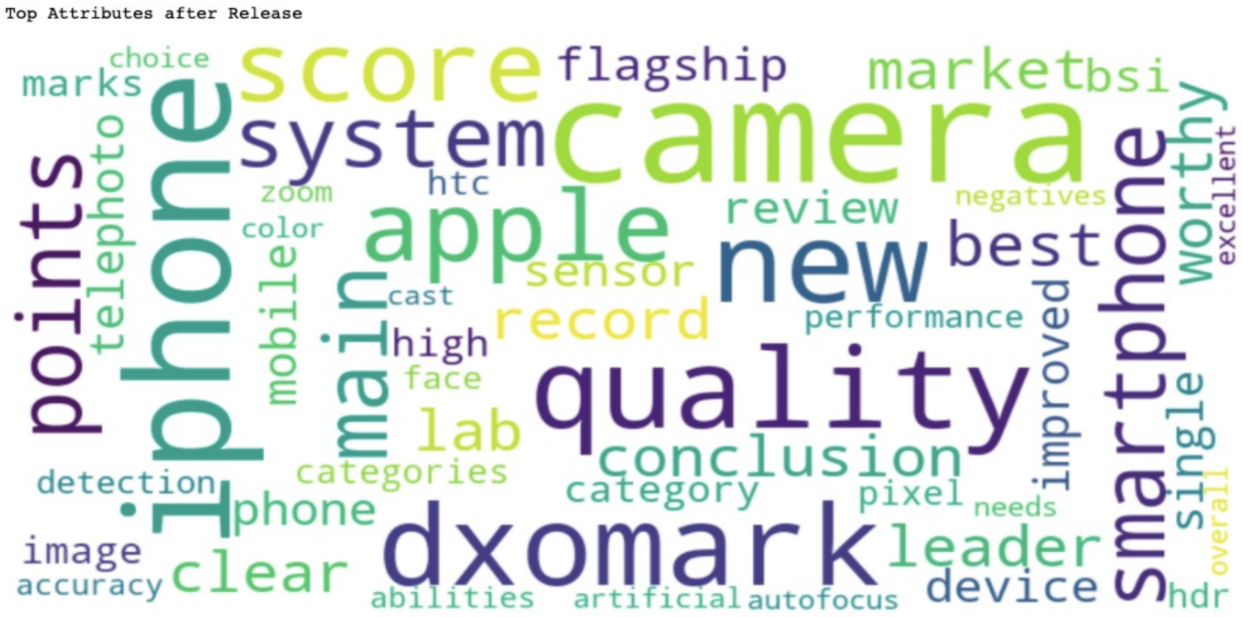
iPhone 8 after release



iPhone X before release

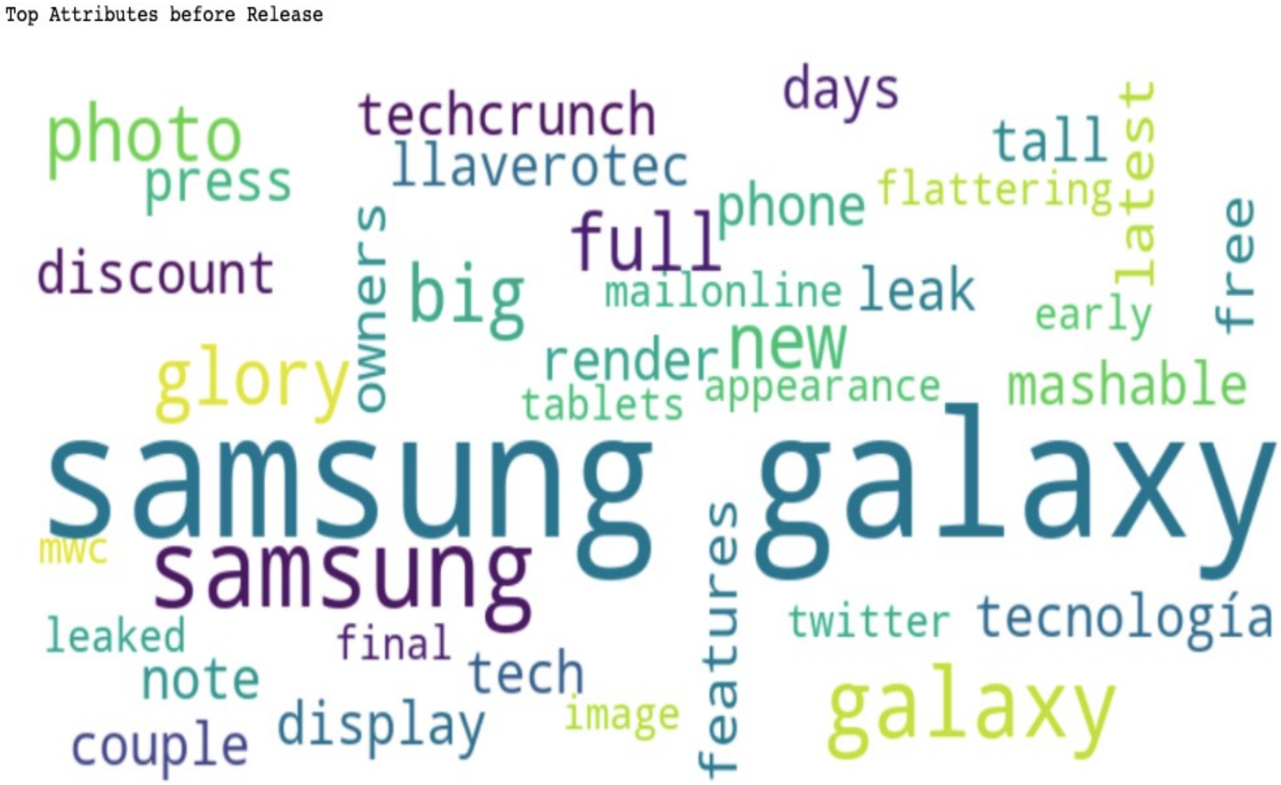


iPhone X after release

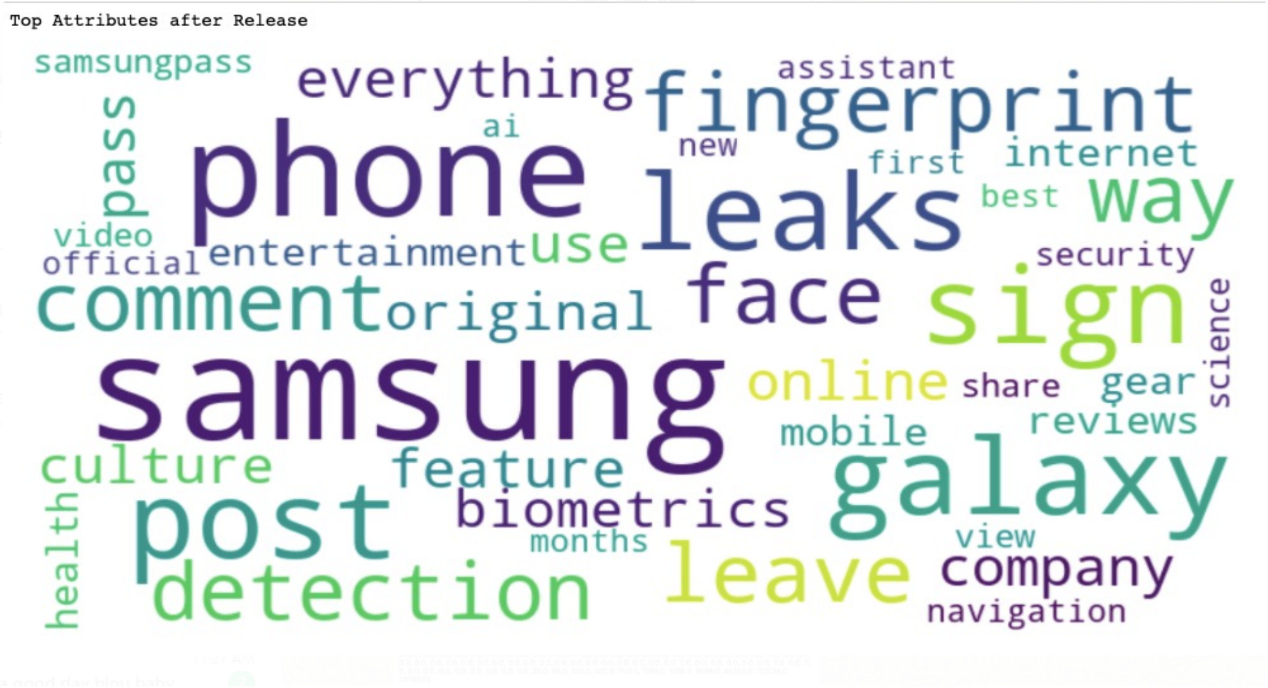


From the word-clouds that we have displayed above for iPhone X, before and after the release, we observe that there seems to be a lot of talk about the camera and its related components of the iPhone X, mainly after its release.

Samsung Galaxy before release

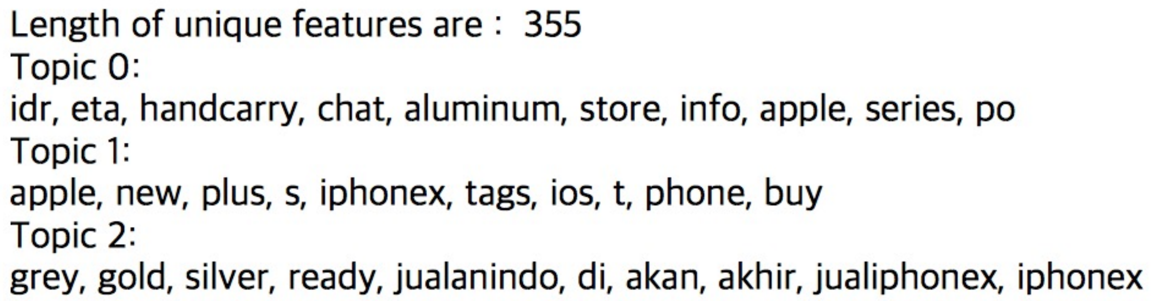


Samsung Galaxy after release

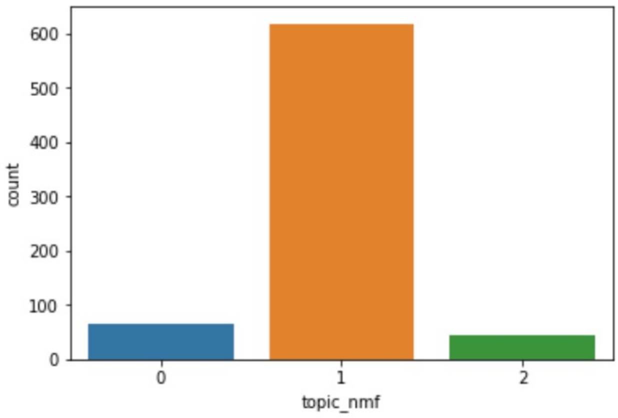


It is obvious that there is not much information to be learnt from this, even though it looks aesthetically pleasing. Thus, to actually figure out what aspects of the phones people are talking about, we need to perform further analysis. We chose to go with NMF Topic Modelling. It is the NLP equivalent of a PCA in Unsupervised Learning methods like K – means clustering. For the purpose of this analysis, we are grouping most frequent words into buckets of 3 and are hoping to gain insight into what features/ attributes of the phones people have been talking about, and if there are any common attributes in a group. The results are interesting, and also act as validations for our process as the grouped buckets predominantly talk about the same kind of feature/attribute. Here are the results for the different phones:

TOPIC MODELLING
iPhone X

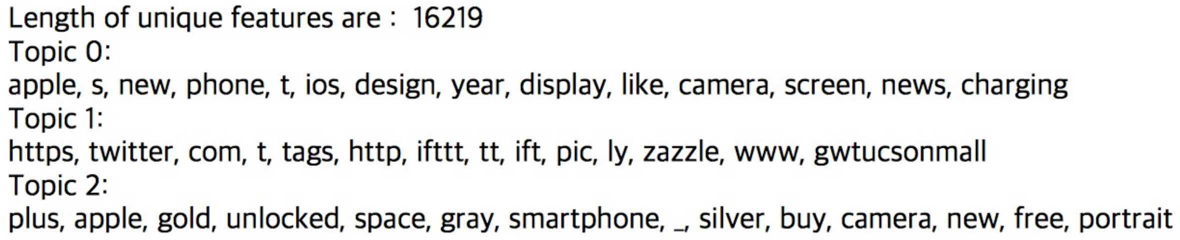


```
Length of unique features are :  355
Topic 0:
idr, eta, handcarry, chat, aluminum, store, info, apple, series, po
Topic 1:
apple, new, plus, s, iphonex, tags, ios, t, phone, buy
Topic 2:
grey, gold, silver, ready, jualanindo, di, akan, akhir, jualiphonex, iphonex
```

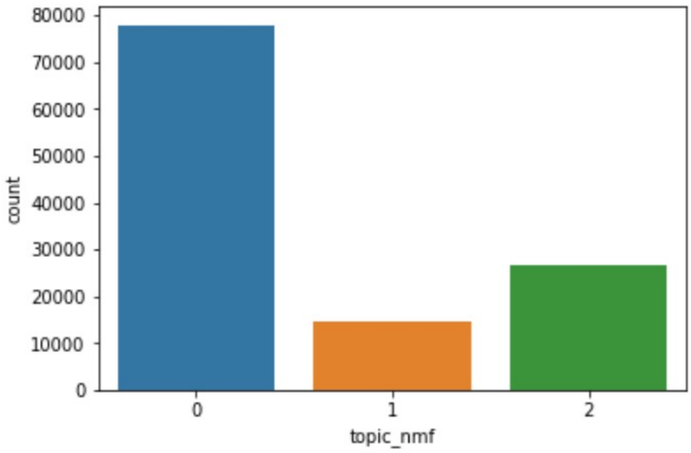


Here, the groups are way more defined and we are able to identify key attributes of the device. The main topic just talks about the phone, and this is what we gathered from the word cloud as well. But Topic 2 deals with attributes related to the design features of the mobile phone. Things like colour, shape form a crux of the conversation people are having about the iPhone X.

iPhone 8



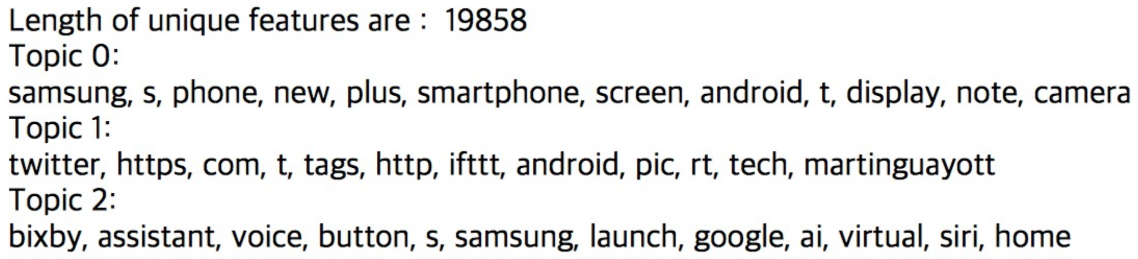
```
Length of unique features are :  16219
Topic 0:
apple, s, new, phone, t, ios, design, year, display, like, camera, screen, news, charging
Topic 1:
https, twitter, com, t, tags, http, ifttt, tt, ift, pic, ly, zazzle, www, gwtucsonmall
Topic 2:
plus, apple, gold, unlocked, space, gray, smartphone, _, silver, buy, camera, new, free, portrait
```



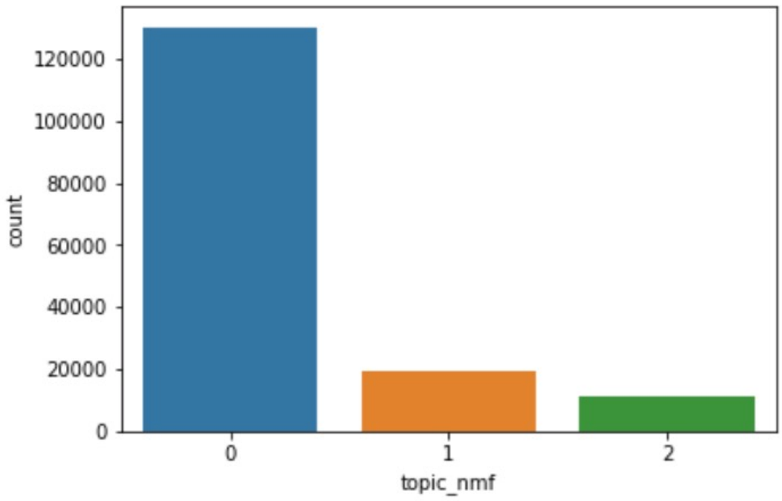
The main topic for the iPhone deals with the features of the phone, and this is validation for the word cloud that we generated. But again, it is interesting to notice Topic 2, which again deals with the aesthetic aspects of the phone. The insights gained for the iPhone 8 is very similar to the one for iPhone X.

As a Product Manager, based on the analysis above, it is not a leap to say that people care more about the design of the iPhones, and not so much about the inherent technology stack. This is not to say that people do not care about the technology, just that people care more about design.

Samsung Galaxy S8



```
Length of unique features are :  19858
Topic 0:
samsung, s, phone, new, plus, smartphone, screen, android, t, display, note, camera
Topic 1:
twitter, https, com, t, tags, http, ifttt, android, pic, rt, tech, martinguayott
Topic 2:
bixby, assistant, voice, button, s, samsung, launch, google, ai, virtual, siri, home
```



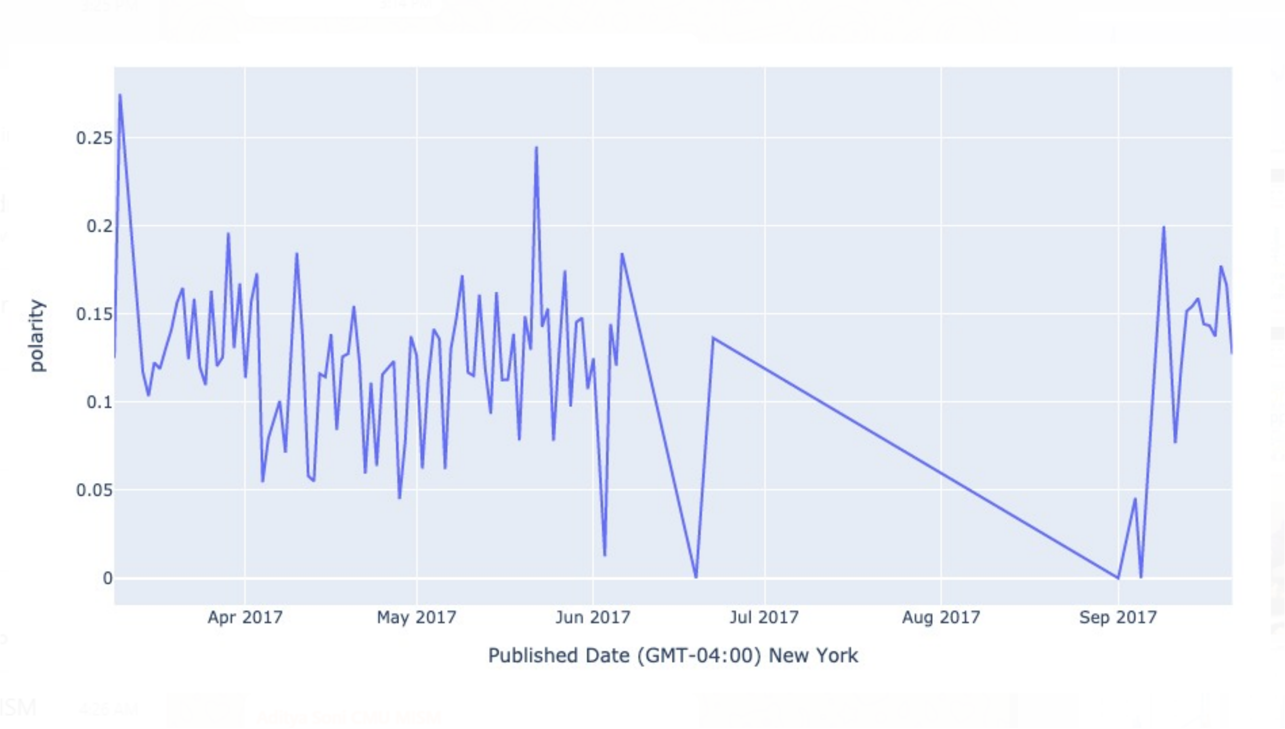
For Samsung, the results are a little bit different. It seems that people care more about the technology stack, more specifically its Artificial Intelligence features.

One possible insight that can be gained from this is that people view Apple as more of a 'Design' company, and view Samsung as a 'Technology' company. Thus, it might make sense for these companies to focus on their strong suits to make products that their customers will buy.

To get a deeper understanding of people's feelings towards the phones, we need to figure out the sentiments towards the products. This can be done in multiple ways, but we are choosing to go with the 'polarity' of every tweet. To do this, we used the TextBlob library that is provided by Facebook. To keep it in simple terms, the library takes a piece of text as input and assigns it a value ranging from -1 to 1. A value of 1 suggests a highly positive emotion and -1 denotes a highly negative emotion. Based on this information, it becomes easier to figure out what people actually feel about the different phone models, and how this changes over time. More specifically, we are splitting the sentiment to before and after the release of the respective products. Here are the results that we found from such an analysis:

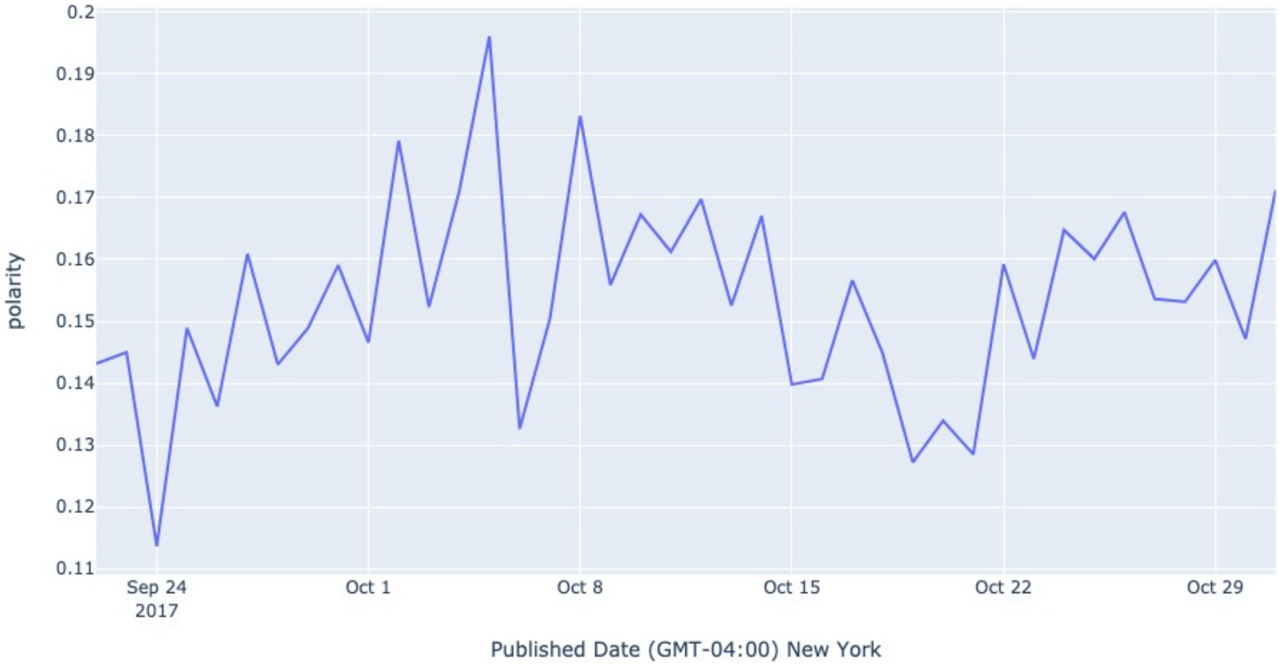
SENTIMENT ANALYSIS

iPhone X before the release of the model



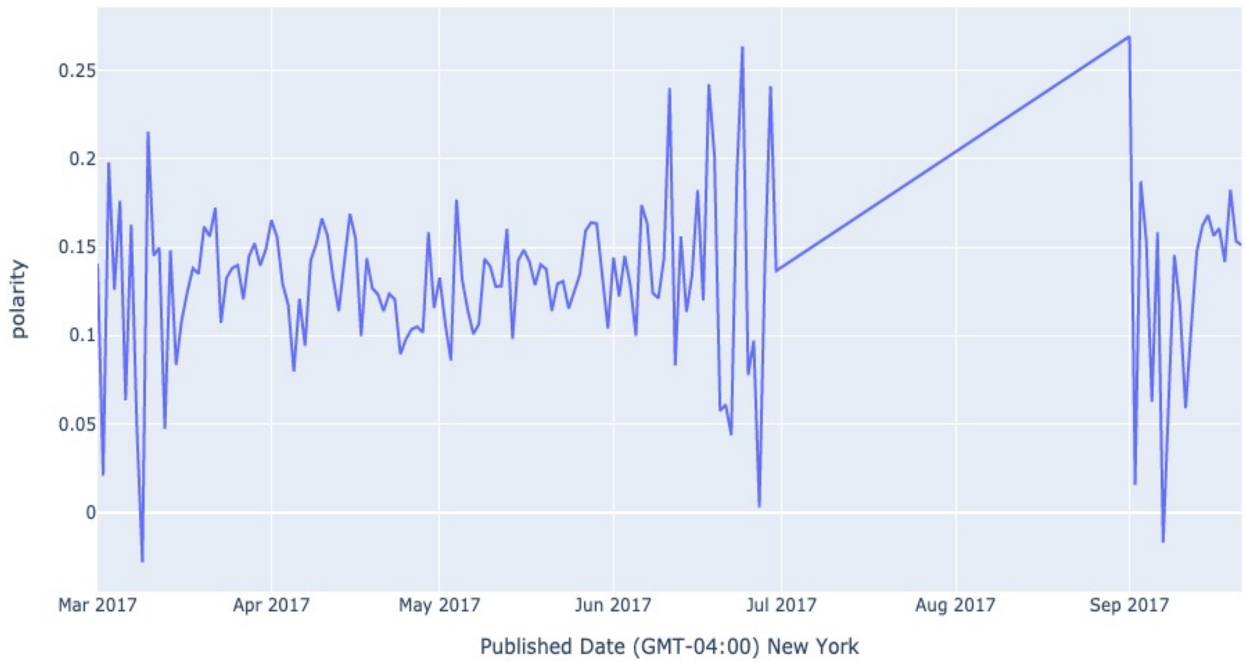
Sentiment Analysis for iPhone X before the release of the model

iPhone X after the release of the model



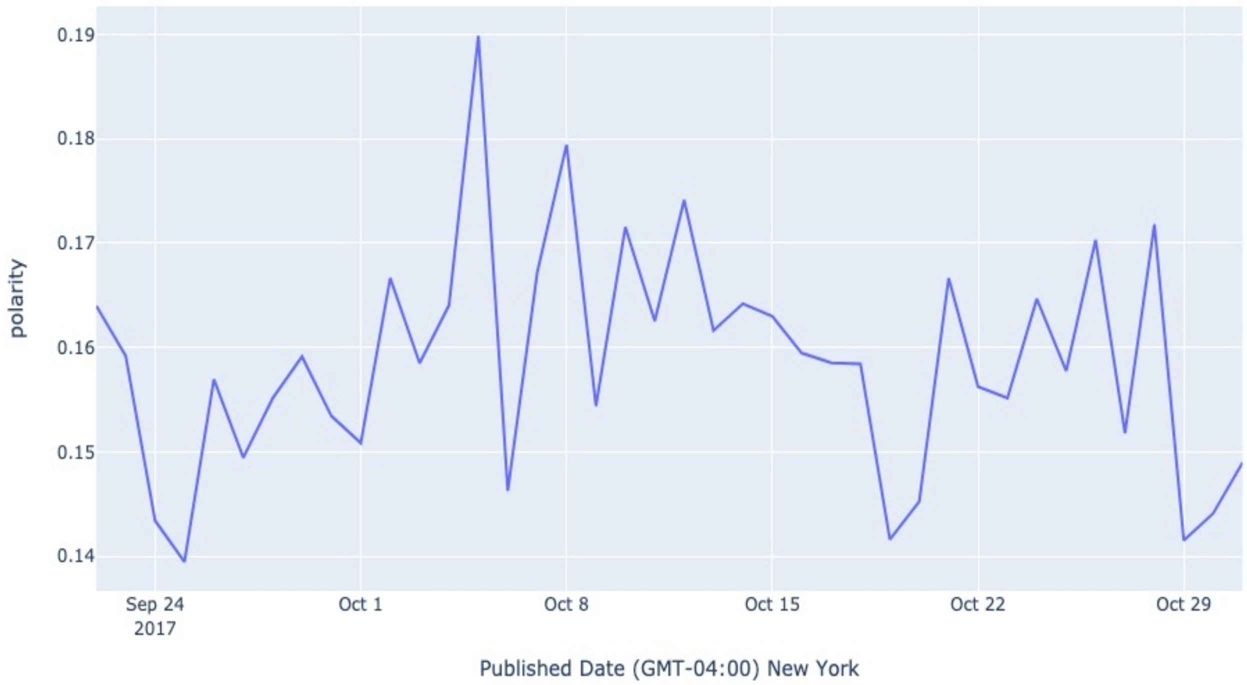
Sentiment Analysis for iPhone X after the release of the model

iPhone 8 before the release of the model



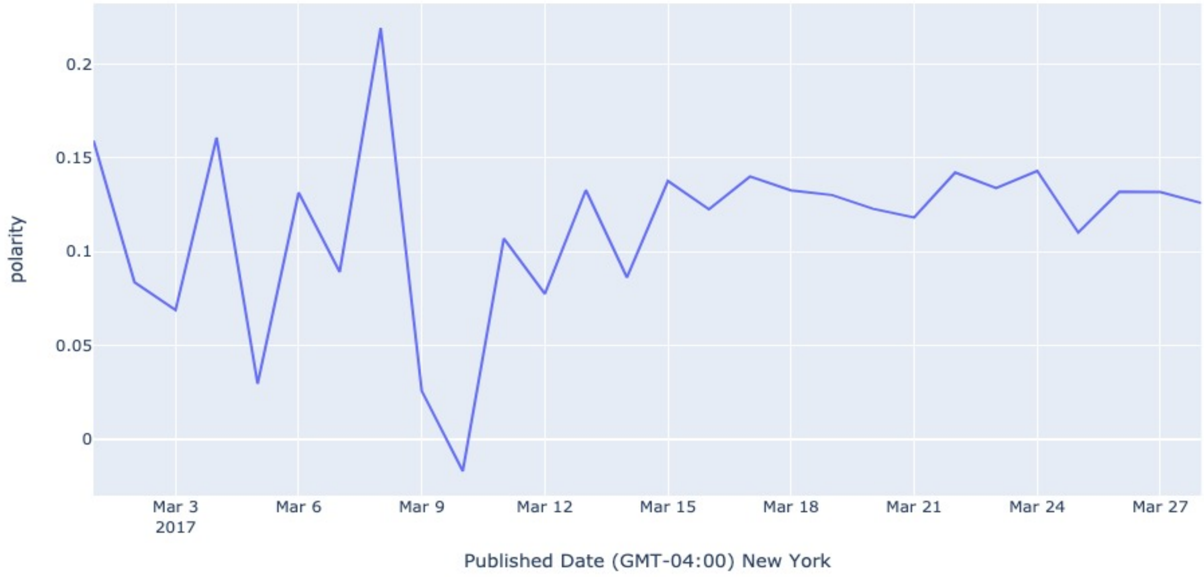
Sentiment Analysis for iPhone 8 before the release of the model

iPhone 8 after the release of the model



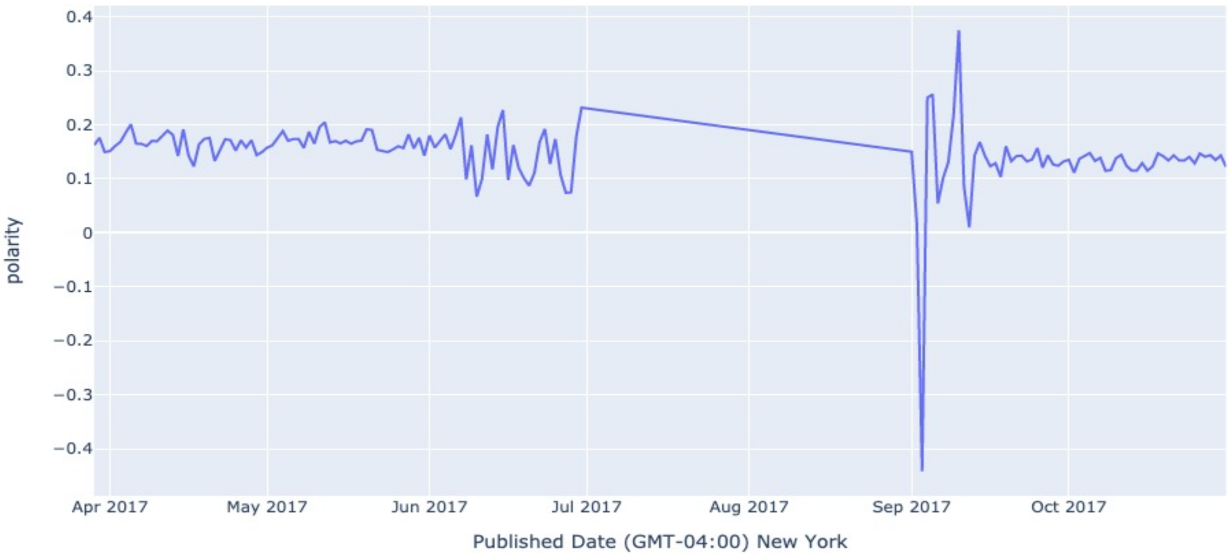
Sentiment Analysis for iPhone 8 before the release of the model

Samsung Galaxy S8 before the release of the model



Sentiment Analysis for Samsung Galaxy S8 before the release of the model

Samsung Galaxy S8 after the release of the model

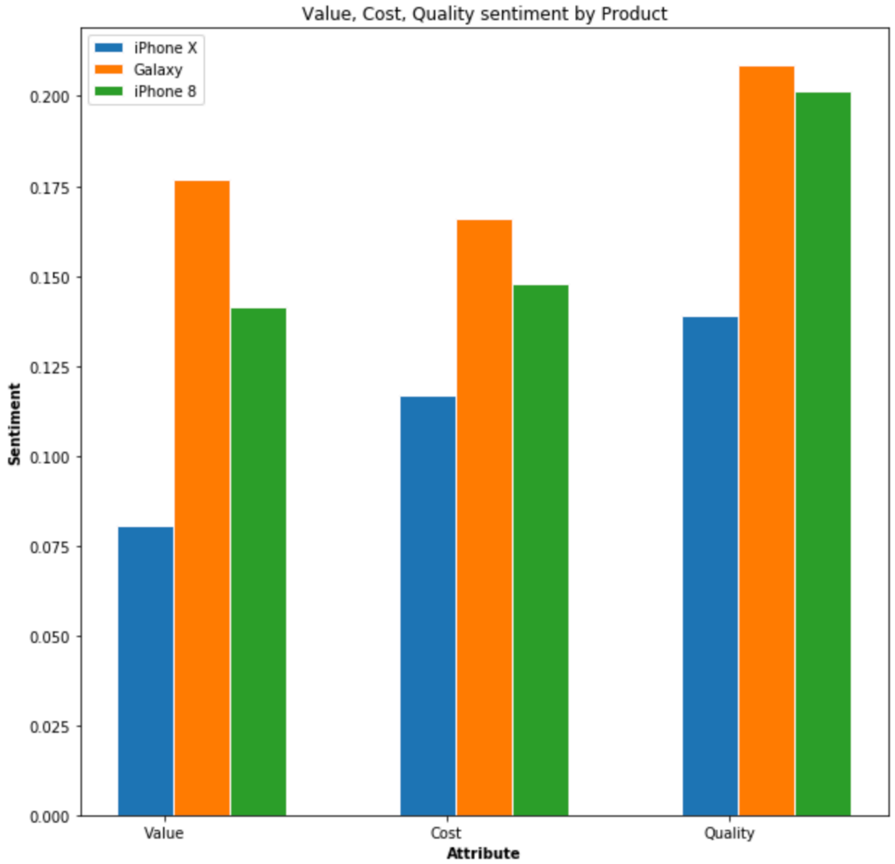


Sentiment Analysis for Samsung Galaxy S8 after the release of the model

We notice a lot of fluctuations in the sentiment, but the most interesting thing to note is the big drop of Samsung's perception during the release of the iPhones. It is such a steep drop, and it must be taken into consideration during future releases. Apart from that, most people have a pretty positive sentiment to all the different models. This could be because both these brands are well established and have a pretty loyal following who know what to expect and have their expectations met regularly.

SENTIMENT ANALYSIS WITH RESPECT TO PRICE, QUALITY AND VALUE OF EACH PRODUCT

In order to compare multiple phones, the 3 most general qualities talked about are Price, Quality, and Value. These might be abstract concepts and vary from person to person, but that is exactly what we want in this analysis as we are looking to find out what the general population thinks about the individual phones. For this task, we are splitting the data based on whether the 'Sound Bite Text' columns contains the words or variations of it mentioned above. After this, we calculated the average polarity in only these rows for each of the 3 phone models, and have compared the values in the following graph:



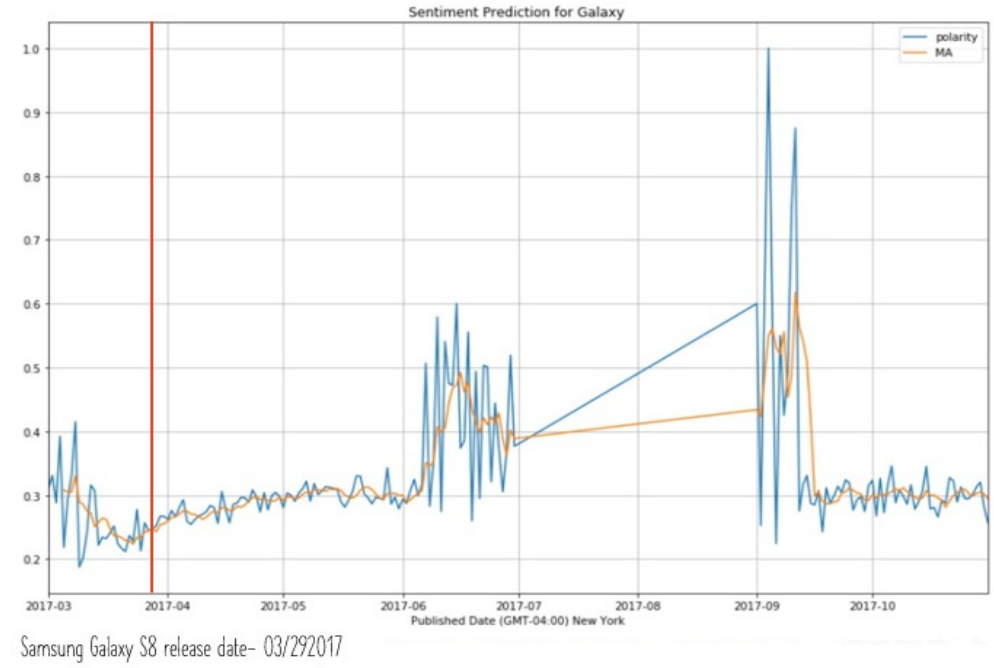
The graph above gives a much deeper insight about what people think about the primary attributes of the 3 different models. Now, the iPhone X does not have a lot of data. This is primarily because people were not aware of its name before its release, and since it released with iPhone 8, a lot of its tweets and sentiments are overlapping with the iPhone 8. But, since the data is ample for Galaxy and iPhone 8, we can draw more legitimate conclusions about these 2 products.

Firstly, people feel that they get more value out of a Galaxy phone than from an iPhone. This can be explained by the sentiments towards the other two features. It must be noted that a more positive sentiment towards price means that it is priced lower. With this in mind, it is no surprise that the iPhone X has the lowest sentiment in this regard, as it was priced at around $1000. The sentiments towards the iPhone 8 and the Galaxy are pretty similar, but the Galaxy edges out here by a small margin. The sentiments towards Quality are where it gets more competitive. There is practically no difference between the sentiments towards the iPhone 8 and the Galaxy, and this is actually a surprise given the historic feeling people have had towards Apple products. But nonetheless, from this graph, it seems obvious that people prefer the Samsung Galaxy to the iPhone models. It should be noted that this analysis might not be entirely accurate. One possible explanation for this could be due to the lack of data that we had. We also had more data for the Galaxy, and that could have added a bit of a bias into the data analysis.

PREDICTION OF ADOPTION

It is interesting to see if we can figure out possible 'adoption' and 'polarity' based on only the sentiment before model release. In the dataset, we do not have a lot information for the iPhone X pre-release due to people not knowing about it and hence we will only be focusing on the iPhone 8 and the Galaxy in this segment. Our approach is applying time-series model using moving average to positive group's polarity with a window of 5 weeks to see if our model predicts the polarity after release well. This is a fairly decent indicator of adoption as it is not possible to accurately figure out if a person has bought a phone based on a piece of text after the release date. Our prediction model based on moving averages gave us the following results:

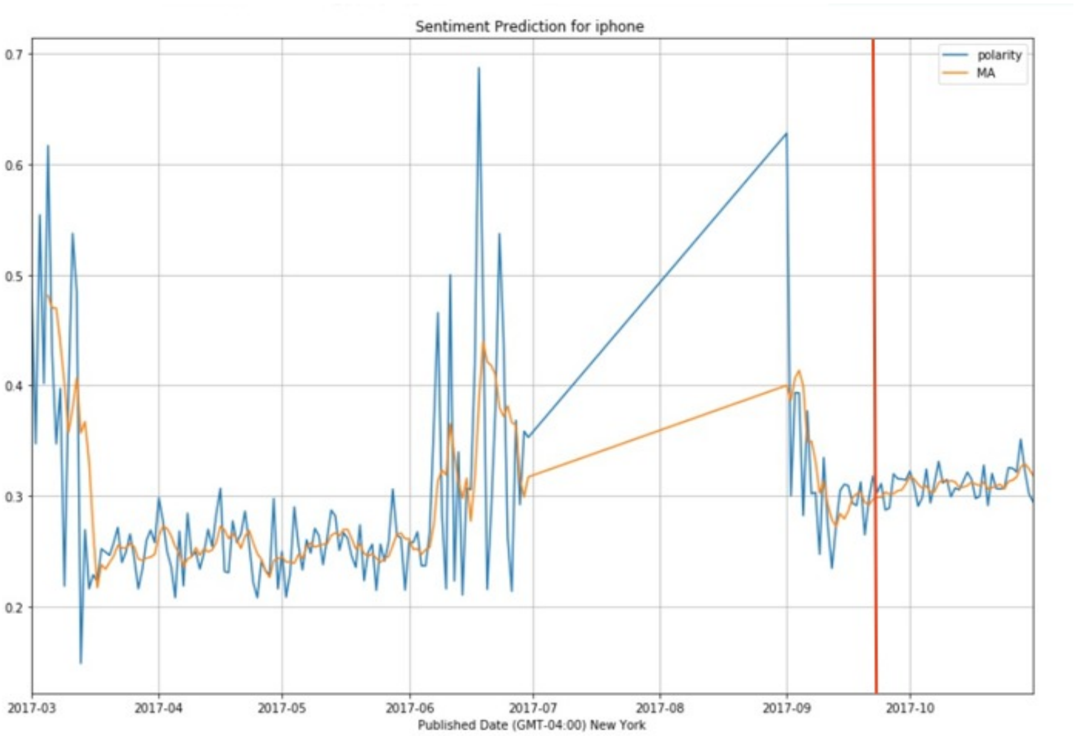
Time Series Prediction for Positive Sentiment of Galaxy After Release



Samsung Galaxy S8 release date- 03/292017

Accuracy of the model - 88.62%

Time Series Prediction for Positive Sentiment of iPhone After Release



iPhone 8/X release date- 09/22/2017

Accuracy of the model - 88.39%

Based on our time series prediction result, we can come up with main common trend for both Galaxy and iPhone model. They both show increase of polarity and it implies that people might be more interested in new products positive features. And based on this trend we can possibly assume that people will be more likely to buy new phone after release. By confirming our prediction rate for both which is about 89%, we consider that Moving Average Prediction is fairly reliable accurate. For the validation method, we use MAE (Mean Absolute Error) to see how our model's prediction is far away we are from the actual values and how generally performed during entire time range.

SUMMARY AND ANALYSIS

From our analysis, it is apparent that both Samsung and Apple are well established brands that compete to gain the lion's share of the smartphone market. Both brands have a well-established set of followers have a fairly decent positive perception towards the brands.

If we had to give a suggestion to these brands, we would suggest Samsung to focus more on the technology aspects of their devices as people seem to be really interested in that. We would also suggest Apple to focus more on their design elements (aesthetics) as they are a brand that prides itself on design and innovation, and this is what people have been talking about in social media.