Google Pixel 5 tweets analysis

Swetha Kallam, Tim Hollinger, Marie-Reine Obama

Feb 4, 2021

# Introduction

In this project, we will be analyzing the tweets about Google Pixel 5 mobile phone. It was released on October 29 2020 in the USA. We will first create a quick word cloud for the data of one month before the release date, just to see what is being talked about the most. Then we will collect data from the end of October till end of January for our analysis. We will perform a brief analysis on the tweets w.r.t time, location, top 10 tweeters, hashtags etc. We will be then be doing a detailed sentiment analysis and topic modeling.

# Before Product Release

Before proceeding with the pre-processing of the data, we will first add some custom stop words specific to our context.

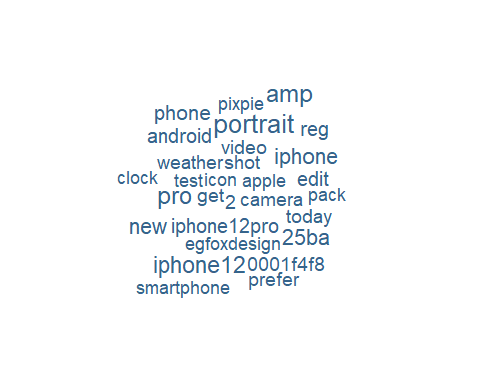
Note that for this quick review, we are not removing any mentions, hashtags etc apart from the obvious ones, because we are interested to know what is being talked about mostly.

## # A tibble: 23 x 2  
## word lexicon  
## <chr> <chr>   
## 1 #pixel5 CUSTOM   
## 2 #pixel CUSTOM   
## 3 @madebygoogle CUSTOM   
## 4 #madebygoogle CUSTOM   
## 5 madebygoogle CUSTOM   
## 6 google CUSTOM   
## 7 pixel CUSTOM   
## 8 pixel5 CUSTOM   
## 9 #google CUSTOM   
## 10 googlepixel5 CUSTOM   
## # ... with 13 more rows

Now we perform some pre-processing steps like

1. Tokenization
2. Converting to lowercase
3. Punctuations removal
4. Removal of URLs
5. Stopwords removal

Now let us see which words are repeated most frequently:



### **Insights**

We see that comparisons are made often with iphone12 , which was also released around the same time. And also the words like portrait, camera, video, shot etc suggest that the camera features were highly talked about and anticipated.

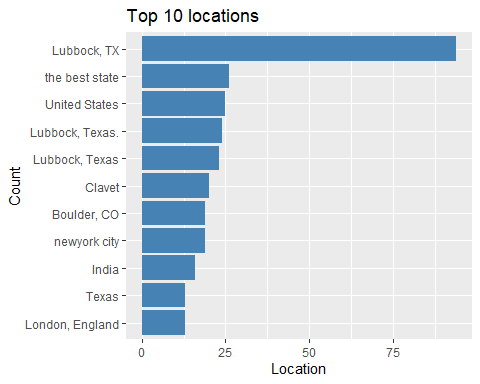
# After product release

We will extract data for 3 months : Nov 2020, Dec 2020 and Jan 2021 for our analysis, 500 tweets per each month. We extract them separately month-wise and then combine them, to ensure that we have an equal distribution.

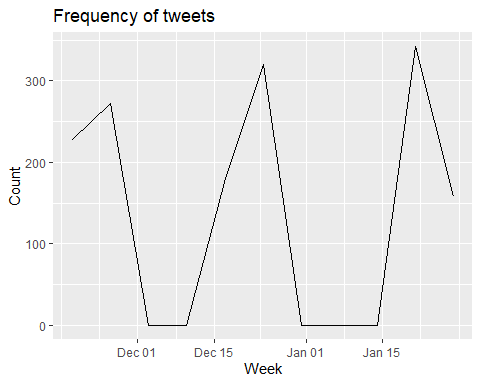
## Performing some quick analysis of tweets

Out of the 1500 tweets we extracted, only few of them has location information. Among those, we are checking what are the top 10 locations from which tweets were posted.

## Selecting by n



Now we will plot the number of tweets per week. Note that the rtweet package extracts tweets in a random manner and this is ofcourse not representative of the entire twitter data collection.



Now we will see the twitter handle (screen name) of the top 10 tweeters.

## Selecting by n

## # A tibble: 14 x 2  
## screen\_name n  
## <chr> <int>  
## 1 thatonehacker5 26  
## 2 LoneStarVarsity 24  
## 3 basketball\_lcp 20  
## 4 stephenlshore 20  
## 5 DaveTaylor 19  
## 6 Lenny\_Bons 19  
## 7 ssaig 18  
## 8 chsenloe 16  
## 9 806hssc 10  
## 10 Alexis\_Cubit 9  
## 11 bestusefultips 9  
## 12 EgFoxDesign 9  
## 13 lcpgirlsbball 9  
## 14 RandyRosetta 9

Below are the top 10 hashtags used most.

## Using `to\_lower = TRUE` with `token = 'tweets'` may not preserve URLs.

## Selecting by n

## # A tibble: 10 x 2  
## hashtag n  
## <chr> <int>  
## 1 #pixel5 1048  
## 2 #teampixel 236  
## 3 #googlepixel 164  
## 4 #googlepixel5 136  
## 5 #google 125  
## 6 #madebygoogle 63  
## 7 #pixel4a 57  
## 8 #nesthubmax 53  
## 9 #paid 40  
## 10 #android 39

Below are the top 10 mentions used most.

## Using `to\_lower = TRUE` with `token = 'tweets'` may not preserve URLs.

## Selecting by n

## # A tibble: 10 x 2  
## mentions n  
## <chr> <int>  
## 1 @madebygoogle 212  
## 2 @pixel5 211  
## 3 @google 192  
## 4 @hubcityprepslbk 125  
## 5 @lcpathletes 123  
## 6 @googleuk 121  
## 7 @lonestarvarsity 120  
## 8 @pchristy11 117  
## 9 @afloradio 112  
## 10 @randyrosetta 93

## Pre-processing:

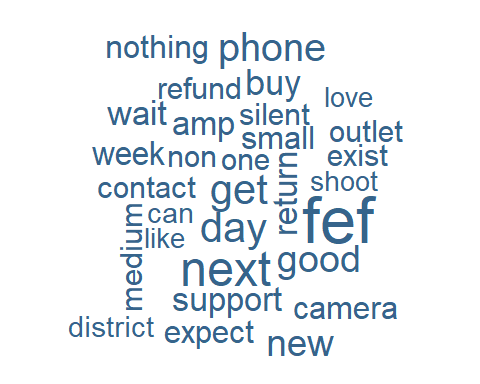
Before the pre-processing, we make sure that words like “shouldn’t” are replaced by “should not” , which will be useful later when we perform sentiment analysis.

Also we remove any hashtags , mentions, URLs, numbers because we cannot attach meaning to those.

We also remove words that are less than 2 characters in length.

Along with the pre-processing steps we have done earlier, we also add another additional step for lemmatization, to convert all words to their root words.

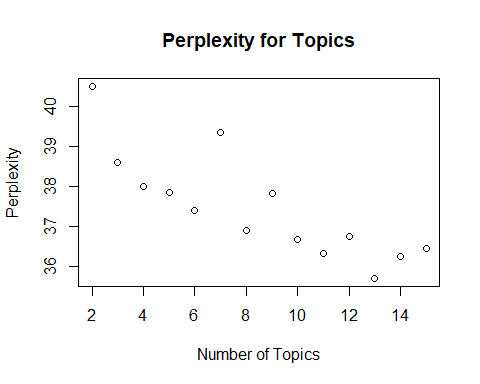
## # A tibble: 3 x 4  
## created\_at text word lemma   
## <chr> <chr> <chr> <chr>   
## 1 2020-11-29 23:38:21 " Oh my lawd, me neither. <U+FF>" lawd lawd   
## 2 2020-11-29 23:38:21 " Oh my lawd, me neither. <U+FF>" neither neither  
## 3 2020-11-29 22:55:40 " Sloppy play thus far. <U+F>" sloppy sloppy

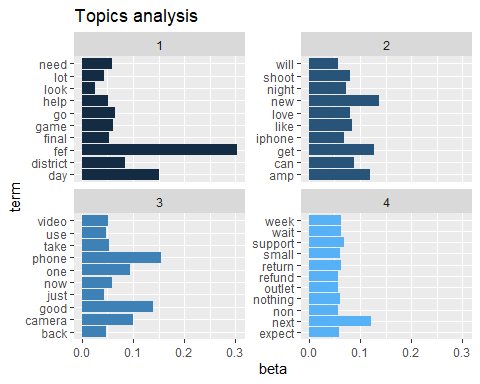
Now let us create a word cloud again to see the most frequent words being used: 

## Topic Modeling

Now we will try to see if we can broadly divide the tweets into different topics. We use the LDA (Latent Dirichlet Allocation) method for this. We create a loop and calculate the perplexity by varying the number of topics from 2 to 15. The lower the perplexity, the better the model, but we also need to keep interpretability in mind.

From the below plot we can see that the decrease in perplexity from 4 to 5 is much lesser compared to reduction from 3 to 4. Hence we stop at 4 and select number of topics as 4 for further analysis.



In the below plot we can see the top 10 words from each topic : 

### **Insights**

Although we cannot make any concrete conclusions from this, we can definitely observe that :

One of the topics is more about the anticipation about the receipt of the product , which is indicated by words like week, wait, expect etc. And it also includes words related to customer complaints like support, refund, small etc.

One of the topics seems to be making comparisons with iphone and its features indicated by words like iphone, amp, shoot, love, like etc.

One topic seems to be talking about camera features among other things.

But the words in few topics are too general to allow for any conclusions.

## Sentiment Analysis

Now let us perform sentiment analysis. The pre-processing for this is almost same, except that we don’t want to remove a few words from our list of stopwords.

Negation words “not, no, nor” change the sentiment of the words following them to the opposite value.

Words like “very, too” alter the value by increasing the effect. Eg: “very angry” is more powerful than “angry”.

Hence we add an extra step to not remove these words from our analysis.

### Afinn dictionary

First we will use the “afinn” dictionary. This assigns a value of -5 to 5 indicating “extremely negative” to “extremely positive” sentiments. We also implement our previously mentioned concept.

If the previous word is a negation word, the value is multiplied by -1 to change its meaning to the opposite.

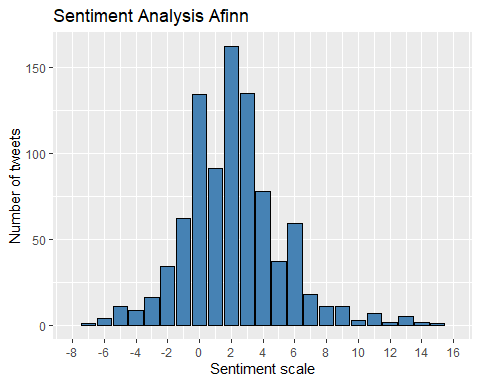
If the previous word is “very” or “too” , we increase the effect of the sentiment by adding 1 to it if positive and subtracting 1 if it is negative.

We then calculate the overall sentiment of each tweet by adding the sentiment values of all the words in the tweet.

The below graphs shows the number of tweets for each value of overall sentiment.

## Warning: Removed 1 rows containing missing values (position\_stack).

## Warning: Removed 1 rows containing missing values (geom\_col).



We can also summarize the sentiments as positive, negative or neutral depending on whether the value is >0, <0 or =0 respectively.

## `summarise()` ungrouping output (override with `.groups` argument)

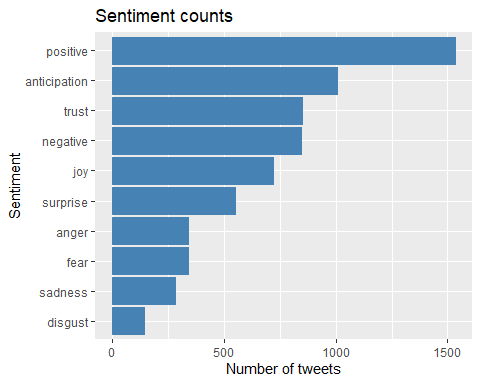
## # A tibble: 3 x 2  
## sentiment\_group num\_of\_tweets  
## <chr> <int>  
## 1 positive 622  
## 2 negative 139  
## 3 neutral 134

### **Insights**

We see that the sentiment is mostly positive. Also the overall sentiment counts extend more on the positive side compared to the negative.

## nrc dictionary

The “nrc” dictionary assigns labels for sentiment to each word.

The below graph shows the number of tweets for each sentiment. 

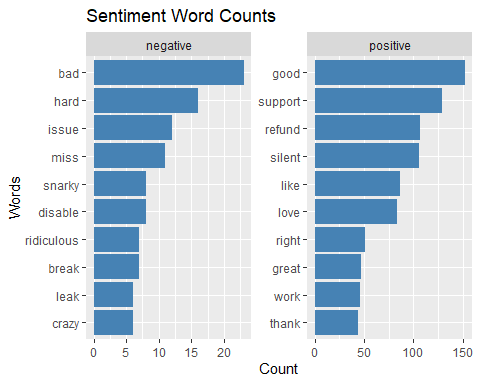
### **Insights**

Apart from “positive”, “anticipation” is the highest emotion , which is of course true because mostly all the tweets are showing excitement about the new release and waiting to acquire it.

## Bing dictionary

Using the bing dictionary, let us explore what are the most used words for both positive and negative sentiment.

The below graphs shows the top 10 words that are most frequently used. However, it is important to note that the counts for negative words are much lesser compared to the positive words.



# Conclusion:

Overall, we have seen that the mobile is often being compared with iphone12. The most talked about feature of the mobile is the camera. We also observed that the overall sentiment is mostly positive.