Sentiment Analysis

Digital Marketing Team Assignment

Fall 1 Term B26:

Ziyi (Summer) Fu

Chunyu (Sarah) Ji

Aastha Thakkar

Ning Xu

Yifan Zhou

October 14, 2019

Contents

Business Understanding

YouTube is a popular video-sharing website and a subsidiary of Google. With over 2 billion monthly users, it is reasonable to deduce that YouTube is generating a mix of positive and negative sentiments.

The company has also experienced its fair share of controversies in 2019. They have ranged from false advertising and unfair copyright claims to cyber bullying and child endangerment.[1] We decided to explore the public sentiment around YouTube, given all the negative media.

Data Cleaning And Processing

We began our analysis by extracting 1200 tweets with the word ‘YouTube’ in them. To clean that data we converted the list of tweets to a dataframe. The information is structured as 16 variables with 1200 observations. The variables include the actual text, favorite count, retweet count, user id, etc.

First, we converted all the text to lowercase so different variations of ‘YouTube’, ’youtube’ or ‘Youtube’ were all treated the same and did not skew the analysis. To ensure the results were not skewed by commonly used, but meaningless words such as ‘I’, ‘the’, ‘and’ we use a ‘stopword’ dictionary and remove all those words.

We also chose to remove all numeric and punctuation information as it did not seem relevant to our analysis. We removed all special characters and emojis because it is hard to analyze within the context of text mining. URLs and whitespaces were also removed for a cleaner dataset.

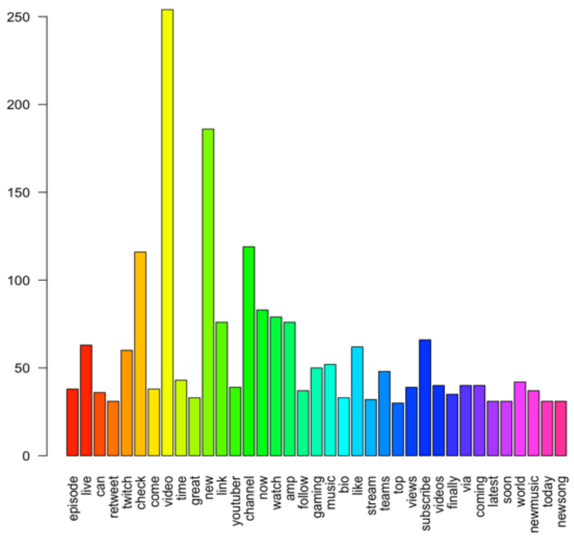
[1] https://www.insider.com/youtube-controversies-that-happened-so-far-in-2019-7#ricegum-and-jake-paul-promoted-a-shady-unboxing-business-for-video-games-to-their-young-fans-1

We then inspect the tweets at each stage of cleaning to ensure that the result is what was intended. After an initial analysis of the most commonly used words remaining in the dataset we, removed ‘youtube’ since tweets were extracted using it, we knew that would appear in every set. We also removed ‘utube’ for a similar reason and removed ‘lets’ ‘gets’ ‘will’ and ‘just’.

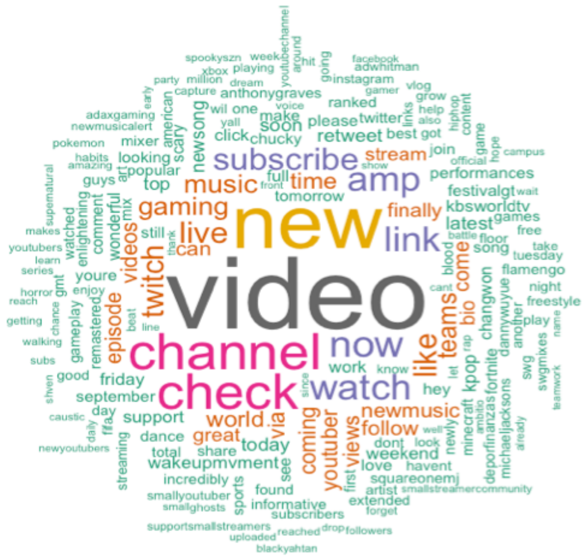
Commonly Used Words

We created a matrix and plotted the most used words by frequency of use. Unsurprisingly, the most used word on YouTube is ‘video’. The next most commonly used word is ‘new’, which is also not surprising given that on average 400 hours of video are uploaded to YouTube every minute.[2] Other words include youtuber, watch, channel, etc. Gaming, stream and twitch are also in the most frequently used words showing the increasing popularity of esports and live streaming. The bar graph shows the top 20 words, while the word cloud shows a bigger range and the size represents frequency.

Graph 1:Top 20 Most Commonly Used Words



Graph 2: Frequency Word Cloud



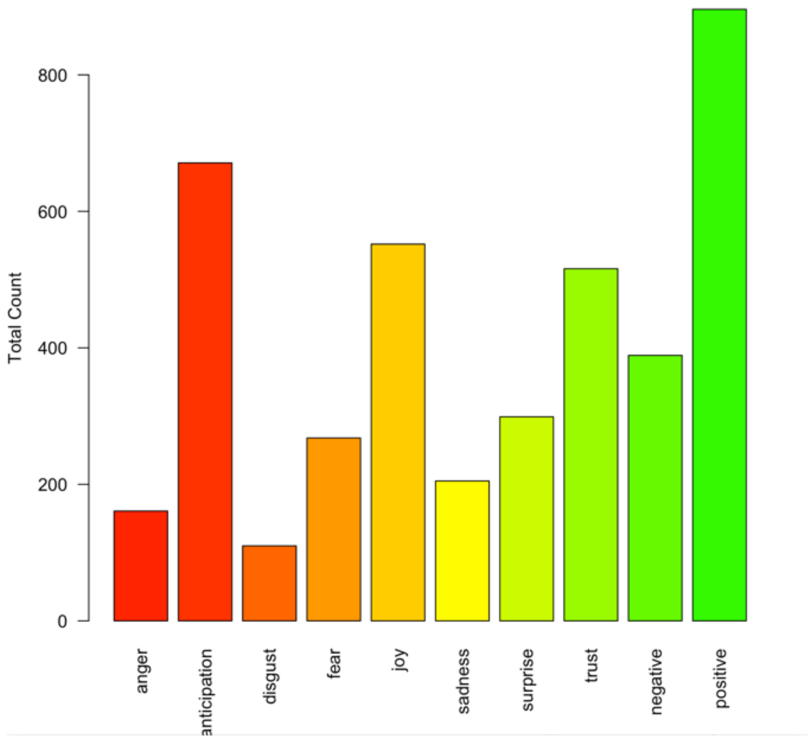
[2] https://www.brandwatch.com/blog/youtube-stats/

Sentiment Analysis

After analyzing commonly used words and trends, we move on to a sentiment analysis to gain a sense of public sentiment in 2019 after a year of controversies.

Words are scored across 10 elements using a sentiment dictionary, which consists of anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative and positive. For instance the word 'finally' has a score of 1 on anticipation, 1 on disgust, 1 on joy, 1 on surprise, 1 on trust, 1 on positive and a 0 on all other factors. An overall sentiment analysis on this sample of 1200 tweets shows that the positive sentiment outweighs the negative. The negative media could simply be a loud minority, not strong enough to harm a strong brand like Youtube. The joy score is higher than fear and disgust combined.

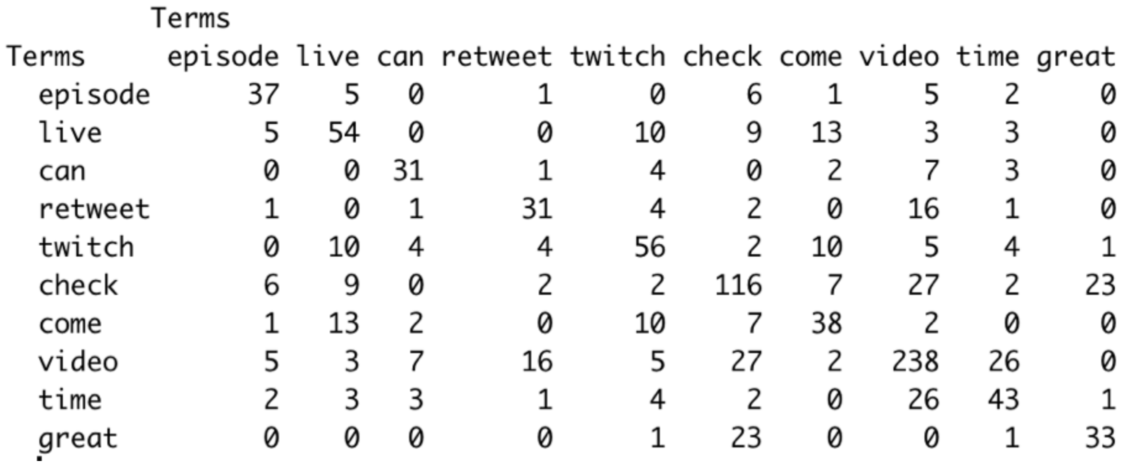
Graph 3: Sentiment Scores for Youtube Tweets



Network Analysis

We then move on to a link analysis to identify different relationships between the various words used in these tweets. To conduct a social network analysis we drop frequency from our term document matrix since we only need to know that a word occurred not how many times. To analyze how many times words appeared together we look at the following tweet adjacency matrix.

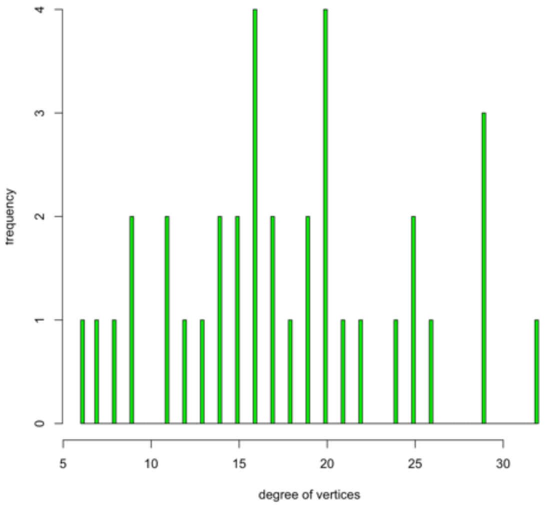
Graph 4: Youtube Tweet Adjacency Matrix



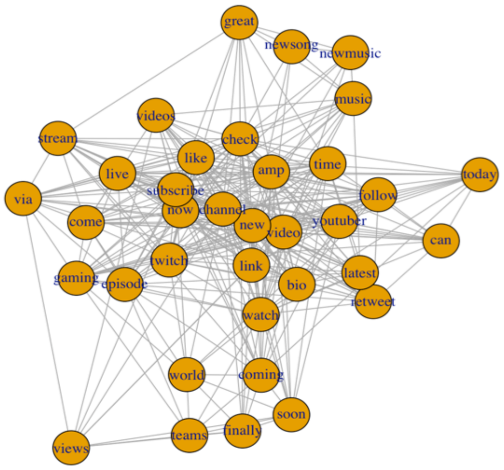
```
         Terms
Terms     episode live can retweet twitch check come video time great
  episode      37    5   0       1      0     6    1     5    2     0
  live          5   54   0       0     10     9   13     3    3     0
  can           0    0  31       1      4     0    2     7    3     0
  retweet       1    0   1      31      4     2    0    16    1     0
  twitch        0   10   4       4     56     2   10     5    4     1
  check         6    9   0       2      2   116    7    27    2    23
  come          1   13   2       0     10     7   38     2    0     0
  video         5    3   7      16      5    27    2   238   26     0
  time          2    3   3       1      4     2    0    26   43     1
  great         0    0   0       0      1    23    0     0    1    33
```

We treat each word like a vertice in the network and assign each word a label of itself. We remove the self-referring terms for simplification. We used an undirected graph to measure these, which means that the connections from words go in both directions. We then compute degree, which is the number of connections between terms. The following is a histogram of the node degrees, which tells us the frequency of the vertices. At the same time, the relationships between words are also visualized through a network diagram.

Graph 5: Distribution of Node Degree



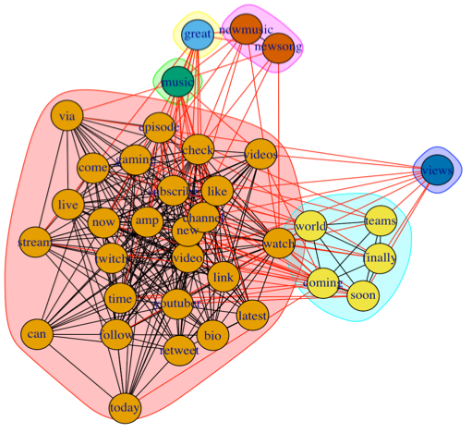
Graph 6: Network Diagram for Words



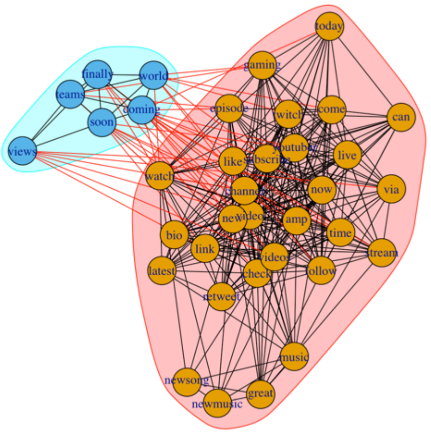
We can also explore communities of vertices through clustering. There are different cluster algorithms, the first diagram is generated through community creation/edge betweenness. The second set of clusters is generated by propagating labels and the third set was created through a greedy algorithm.

The first method generates lots of small clusters, there is a cluster for the word 'music' alone and another one for 'views' alone, while many other words are clumped together. This takes away from the readability. The second set of clusters is just 2 very broad categories.

Graph 7: Clustering via Edge Betweenness

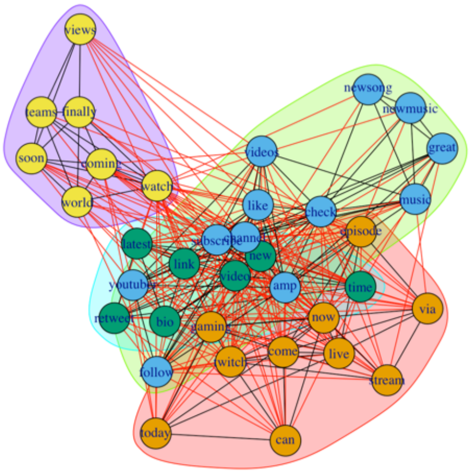


Graph 8: Clustering via Label Propagation

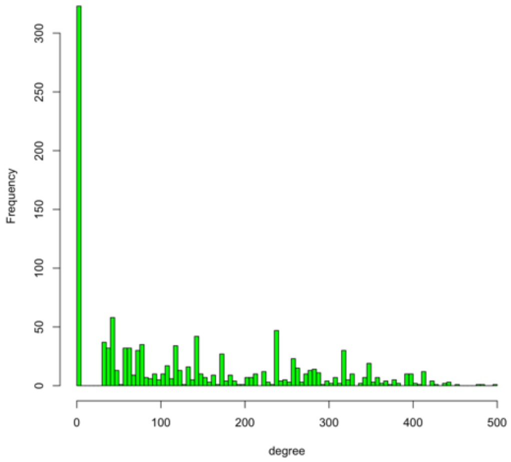


The third set seems more informative with 4 groups. The orange nodes seem to be gaming focused words. The blue nodes seem to talk more about music. The yellow nodes are focusing on new video launches with words like ‘coming’ and ‘soon’, which would occur together frequently. The green node is mostly likely Twitter users promoting their Youtube videos with ‘link’ in ‘bio’.

Graph 9: Clustering via Greedy Algorithm

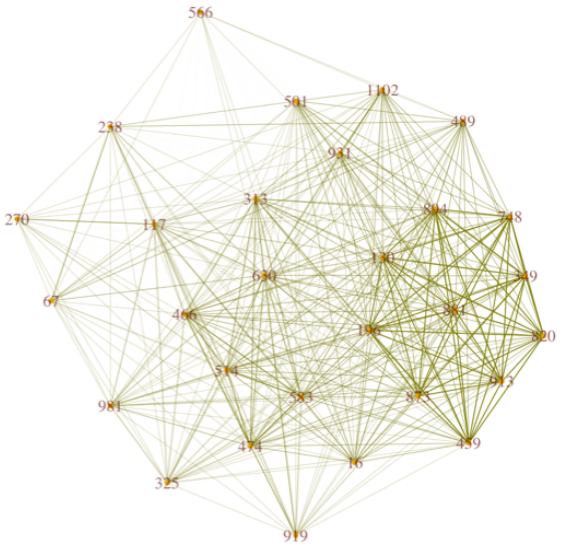


Graph 10: Histogram of Degree



A similar link analysis can also be conducted using tweet IDs as vertices. 0 is the most commonly occurring frequency, which means a bulk of tweets are not connected to any other tweets.

Graph 10: Diagram of Degree



Tweets that have very few connections are the ones on the outside of the diagram such as tweet 566, 270, 919, etc. These tweets include content such as “Started a YouTube video &amp; said \"yeah I can eat to this\" \n#youtube #YouTuber #smallyoutuber #follow #Subscribe https://t.co/c5zgP6UQKC” This does not make too much sense and is not very popular.

Conversely tweet 630 and 130 with content like “RT @New_YouTubers: NEW #YOUTUBE VIDEO\n\nIntroduce Your Channel | Tuesday October 08 \n[CLICK THE LINK &amp; INTRODUCE YOUR CHANNEL IN THE VIDEOS” This seems relevant and connected for many YouTube users so it is in the middle of the diagram.

Conclusion

All in all, we used several visualization methods and network analysis to analyze the 1200 tweets. These help us conduct contextual mining of text which identifies and extract subjective information in source material, and helping YouTube to understand the social sentiment of their brand. This can also help Youtube market to their audience based on clusters. There are clusters about gaming, music, creating content, and video promotions. Youtube can customize homepages based on these themes. They can also show users ads based on the words they are using and analyze the degree to which a user can be an opinion leader or influence.