`YouTube/Twitter Trending Analysis

EECS 4415 Final Project Proposal

Andres Rojas
Lassonde
York University
Toronto Ontario Canada
andres15 @my.yorku.ca

Juyoung Kim
Lassonde
York University
Toronto Ontario Canada
jyk3216@my.yorku.ca

Adrian Winkler
Lassonde
York University
Toronto Ontario Canada
adrianw@my.yorku.ca

Richmond Truong
Lassonde
York University
Toronto Ontario Canada
richt@my.yorku.ca

ABSTRACT

Through the rising popularity of social media, low cost video production, and viral marketing, YouTube has become a monolith in the cyberspace. A website where anyone can post videos, without needing preliminary training or skill and a trending tag can put you on a stage with other popular videos and creators in your region. We have two main components to our project, batch data analysis and cross platform streaming analysis. Using a batch dataset that was scrapped from YouTube trending for multiple different regions. Some of the attributes this dataset hasare the following: trending date, channel, video title, category, tags, thumbnail, description, views, comments, likes and others. We identified a few of these attributes in depth that we would like to explore. We identified the video title, tags, description and thumbnail as the most important factors for becoming trending. Using this batch dataset, we performed sentimental analysis on the description, video title and color analysis on thumbnails. The second component consisted of performing cross platform streaming analysis. We setup a streaming platform that would find currently trending videos on YouTube and check how many people tweet about those videos on Twitter. Using the Twitter API[1] (tweepy) we gathered a fraction of the tweets about the YouTube trending videos in real time in combination with out batch dataset. Using this data we can develop an deep intuition regarding the relationship between a video trending status on YouTube and its impact on the Twitter community.

KEYWORDS

YouTube, Trending, Viral Media, Twitter

Introduction & Motivation

YouTube is a video platform that host content from millions of creators and billions of viewers[2]. When viewers watch a video from a creator; ads are shown before or during the video and YouTube makes money. In exchange YouTube pays a small portion of its profits to the content creator.[3]

Twitter is a platform used by millions of users. Twitter's main features are short blurbs of text (less than 160 characters) called tweets. Twitter has become a fast paced platform where news, ads, and user created content to spread to a worldwide audience. Due to the nature of social media, most major YouTube creators have profiles on many different platforms, including Twitter. Often, to garner more attention and views, content creators will alert their twitter followers when they are about to release a project. Both of these platforms are driven by real time data and create terabytes of information each minute.

YouTube pays its content creators based on the amount of views that the creator generates to the YouTube platform. Most notably the content creator know as "pewdiepie" made $12 million USD in the year of 2017[4] by making weekly videos. Many people have found successful careers out of posting videos online, commonly referred to as Youtubers. Due to the lucrative nature of being a YouTuber there is lots of competition. This makes it difficult for smaller channels to make successful content. Trending videos can generate tens of thousands of dollars for a YouTuber, thus trying to optimize every factor is highly in their best interest. Our analysis sheds some light on how to target your video to each of the categories that appear on YouTube trending.

In addition, a big portion of making a video hit trending is to market is properly on other social media platforms like

[1] http://www.tweepy.org

[2] https://www.youtube.com/intl/en/yt/about/press/

[3] https://support.google.com/youtube/answer/72857?hl=en

[4] https://www.forbes.com/sites/maddieberg/2016/12/05/the-highest-paid-youtube-stars-2016-pewdiepie-remains-no-1-with-15-million/#7909cdd17713

ACM Reference format:

Andres Rojas, Juyoung Kim, Richmond Truong and Adrian Winkler. 2018. YouTube/Twitter Trending analysis:EECS 4415 Final Project.

Facebook, and Twitter. In our analysis we look at how often a trending YouTube video URL is contained in a tweet. Our combination of Batch analysis and Cross platform streaming analysis aims to find the patterns within successful trending content.

1 Data

1.1 Batch

The batch data we acquired was a dataset of 360 MB containing trending data for 5 different regions over the course of 9 months (nov 14 2017 to june 14 2018). This data contained detailed information for over 40k trending videos including video id, tags, thumbnails, number of view, etc. We also wrote scripts to gather current YouTube data, updated YouTube data from the dataset, and non streamed twitter data.

1.2 Stream

Alonge side the batch data we analyzed, we also collected real-time data from both YouTube's trending section and Twitter. Since YouTube updates its trending page every 15 minutes[5], we scheduled a scraper to follow this same cycle. Whenever a new video was promoted to the trending page, the Twitter stream would be updated to track the id of said video. Depending on the type of video that is being tracked, the Twitter data we device may be sparse. This gives us a look at what kinds of videos will have the most impact once they become trending.

2 Data Analysis

2.1 Thumbnail analysis

Thumbnail analysis was done on the batch dataset of YouTube trending downloaded from Kaggle containing YouTube trending videos from November 14, 2017 to June 14, 2018. The dataset provided URLs to where we could download the image, they were not a part of the dataset. The actual analysis was done using a MapReduce style to allow for scale ability. The dataset had not been cleaned and required some cleaning to be done before we extracted the URLs. The Mapper python script downloaded the image from the source and perform RGB analysis on the image, the output was a color ("lightgrey", "limegreen", "skyblue") followed a count. Some of the issues encountered were that the URLs was broken or that the image thumbnail had been replaced by a default image. To denote this, a separate tag "failed" was created to replace the color. Next the output of the Mapper was fed into a Reducer where it summed the results of each color. This analysis will show the most prominent and successful thumbnail background choices.

2.2 Sentiment analysis

Sentiment analysis was done on the batch dataset of YouTube trending downloaded from Kaggle containing YouTube trending videos from November 14, 2017 to June 14, 2018 for Canada, France, Germany, United States, and United Kingdom. The sentiment analysis was done individually for each country listed but graphs were done for only Canada.

The sentiment analysis was done using MapReduce to process and compute quickly. MapReduce was chosen to process this data because is very scalable and therefore if our dataset was increased by multiple folds the process time would not increase as much. First we used a Mapper that did the following: read the csv file to get the descriptions and titles of the videos, used textblob module to get the sentiment of the titles and descriptions as 1,0 or -1, then outputted category number title sentiment, description sentiment, and video id. Then it was sorted and sent to the Reducer. The data was then sorted by category number and sent to two different Reducers. One summed all the values while the other game the results as percentages. The sum Reducer would get all the summed values for each type (positive, negative, and neutral) for each group (title and description). It would calculate the average sumed sentiment (positive - negative) and output that data with the delimiter ":". The percentage did the same except the values were divided by the total number of values for example titles positive would be positive count / (positive + negative + neutral counts). This way we get to see what the sentiment was overall without being affected by amounts. Then each of the two outputs had their category numbers replaced by the category names.

The data was graphed to give the results seen in figures 5.2.1 to 5.2.4. The percentage graphs values were input rounded down to 4 decimal places which is more accurate than the eye can see on the graph to tell the difference.

2.3 Top categories

Since we had data for tending videos over 5 different regions, we wanted to see what category was most popular in each region. This was done through the efficiency and simplicity of the map reduce algorithm. This gave us a list of the number of times each category appeared in the trending page for each region, this includes songs that lasted for multiple day on the trending. From here, the top three were selected for each region.

2.4 Real time

Once we had a stream set up to Twitter and a scrapper collecting YouTube trending data, the next step was to derive some sort of intuition from this data. Since the data is separated into 15 minute chunks, we wanted to see how the number of tweets pertaining to a specific YouTube video changed overtime. Within the Twitter stream we extracted the uncompressed URLs in the tweet and sent it over to the spark client.

The spark client is responsible for determining which URLs pertaining to each YouTube video, any other link is considered noise and is ignored. The URLs that do belong to YouTube videos are then mapped by the spark client. This will then be reduced by a simple aggregation function to get the number of occurrences of each video id across all tweets we see. These are recorded and sent over to a visualization program. Similar to the scheduling system in the Twitter stream client, these values are also reset every

[5] https://support.google.com/YouTube/answer/7239739?hl=en

15 minutes. This allows us to synchronize our stream, analysis, and visualization platforms together.

The processed data arrives at the visualization program which displays real-time analytics on a line graph. The x-axis represents time and the y-axis represents the amount of tweets that happen in that time interval. Each video has its own line.

3 System Architecture

3.1 Batch Processing

For batch processing, the sentiment analysis was done by running on a MapReduce in a single pipe. The data was not real time but instead downloaded from Kaggle. The choice to use only a single pipe was chosen due to the small size of data that is only less than 50,000 entries for each country which a Mapper on a single machine is able to process under a second. For datasets magnitudes larger it may be more plausible to use a Hadoop system or any other large scale MapReduce platforms to run the MapReduce. The Reducer had to process less than the Mapper and therefore runs even faster due to the less computation required and therefore less incentive to use a large MapReduce platform. The MapReduce for this analysis was run by running the Mapper and piping it to a sort then into a Reducer. Then any category number was replaced by a category name. Since the data was offline it was hard coded into the html files for display and visible through a browser.

3.2 Real Time

The architecture we chose to implement for our stream data had to allow for real-time and the possibility to include batch analysis later on. We ended up implementing a sort of sudo-lambda architecture which allowed for data to be analyzed in real time as well saving data for later projects. A speed layer was implemented in the form of an apache spark client which both read and processed data before pushing it onto a visualization system. At the same time that the real time data is being fed into the spark client, a batch layer is responsible for dumping data into our data warehouse. Using the ETL principle, the data is extracted from the YouTube trending page and converted into a dataset with a well defined schema. This date can then accessed later to extend the original dataset we got from Kaggle.

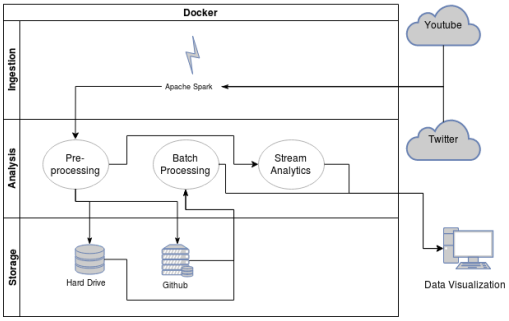

Figure 3.2.1: Analytics Platform Architecture

4 Evaluation and Results

Using an acquired dataset on trending YouTube videos we did color analysis on thumbnails, and sentiment analysis on descriptions and titles. We also gather data real time with the YouTube and Twitter apis. With that data we see how many tweets happen for each particular YouTube video in the trending section of YouTube.

4.1 Thumbnail Results

The color analysis of the trending YouTube video produced unexpected results. Within the YouTube content creator community the typical advice given to creators for thumbnails is to make them colorful. However, our analytics showed that more than 50% for the thumbnails most common color were Black or Grey with blue being the third most common background color as shown in Figure 4.1.1. Some of the reasons why this might be the case is that the black, grey background give a spotlight effect and bring the users attention to the foreground. This insight can help YouTubers better target their content toward audiences.

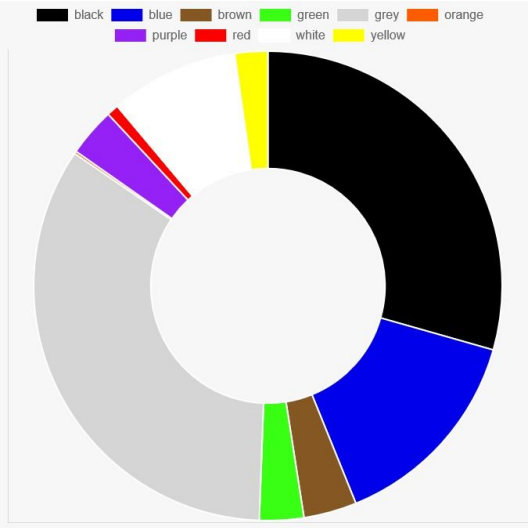

Figure 4.1.1 - Breakdown of most common color used in thumbnail. Black: 29%, Blue: 15%, Brown: 4%, Green: 3%, Grey: 33%, Orange: ~0%, Purple: 4%, Red: 1%, White: 9%, Yellow: 2%.

4.2 Sentiment analysis results

The sentiment analysis on the trending YouTube videos brought unexpected results. The dataset used was the trending YouTube videos in Canada. We had hypothesised that the titles of the videos would be negative while the descriptions would be either positive or neutral. We hypothesised this because of the current state of the globe. Also we believe that negative emotions are easier to invoke and cause a large discussion over. In figure 4.2.1 and 4.2.2 the titles of the trending videos were overwhelmingly neutral with the exception of Movies and YouTube Movie Trailers. Canada did not seems to have YouTube Movie Trailers category appear at all in trending while the Movies category

was overwhelmingly positive in the sentiment analysis for the titles.

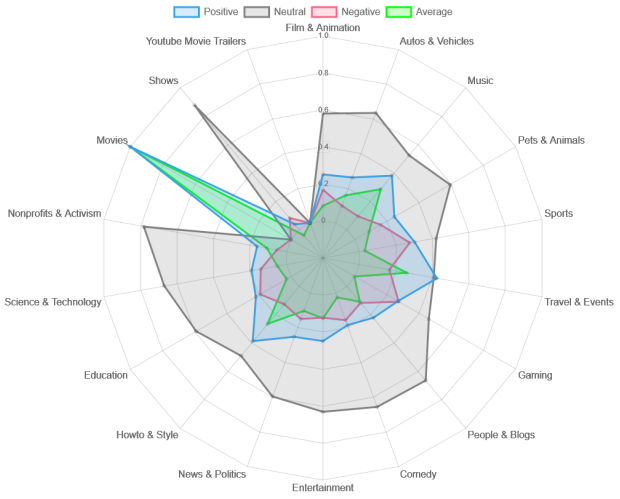

Figure 4.2.1 Title analysis percentage

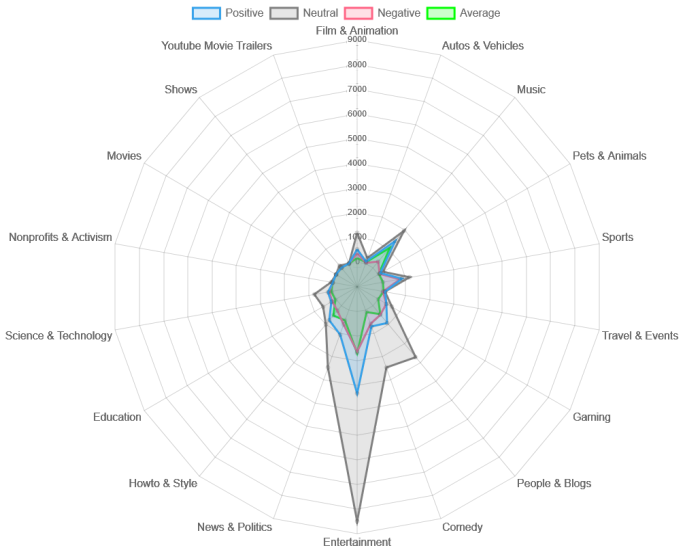

Figure 4.2.2 Title analysis sums

The descriptions of the trending videos on YouTube tell a different result. Majority of the videos had positive descriptions with the exception of Non-Profits and Activism where it is positive but not overwhelmingly. This is clearly shown on figure 4.2.3 where the averages (green web) being over 50% in the positive for all categories except the two mentioned categories. The averages also show how overwhelming the positive the descriptions on trending are because it nearly fills the entire area of the positive values which nearly covers everything else.

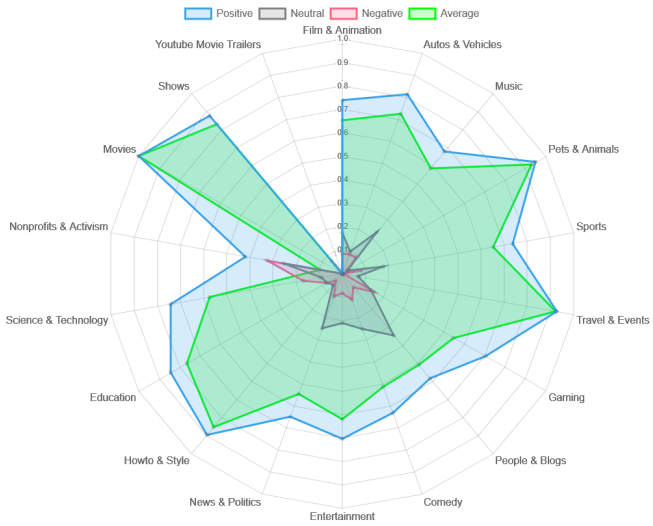

Figure 4.2.3 Description analysis percentage

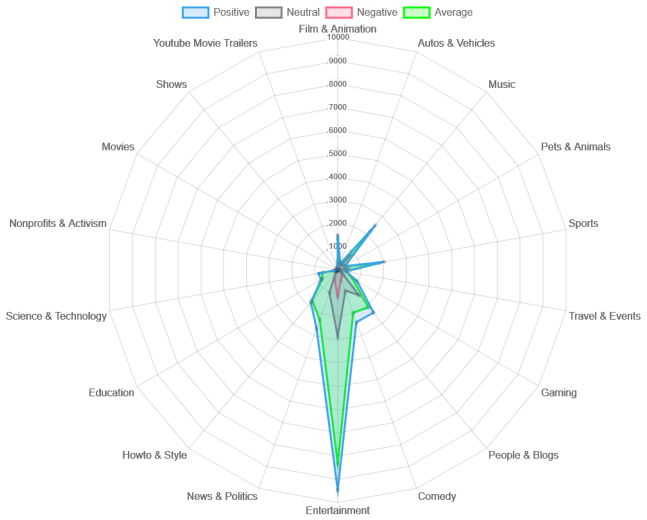

Figure 4.2.4 Description analysis sums

Most of the videos on trending were entertainment by an overwhelming amount as seen in figures 4.2.2 and 4.2.4. The descriptions were overwhelming positive and the titles were mostly neutral. From our data we can say that a video that is in the entertainment category with a neutral title and positive description is more likely to be in the trending of YouTube.

4.3 Most trending category per region

After running our analysis of the top categories we found in our data based on region we found a few interesting results. For one, Entertainment was the most common category found in every region except for United Kingdom.

Region	1st place	2nd place	3rd place
CANADA	Entertainment	News & Politics	People & Blogs
GERMANY	Entertainment	People & Blogs	News & Politics
FRANCE	Entertainment	People & Blogs	Comedy
UNITED KINGDOM	Music	Entertainment	People & Blogs
UNITED STATES	Entertainment	Music	Howto & Style

This could be due to a couple of factors, such as the entertainment category not being well defined (covering a broad range of videos) or videos that focus on being entertaining are shared more often.

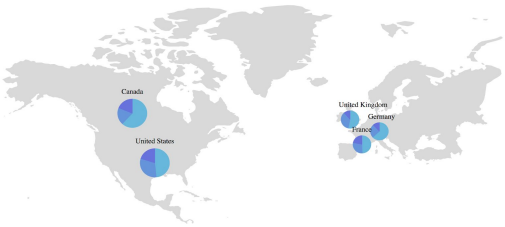

Figure 4.2.1 Regional map. Each pie chart represents the distribution between the top 3 categories per region

An interactive version of this graph can be found under *Top3CategoriesMap.html*

4.4 Real time Twitter and YouTube

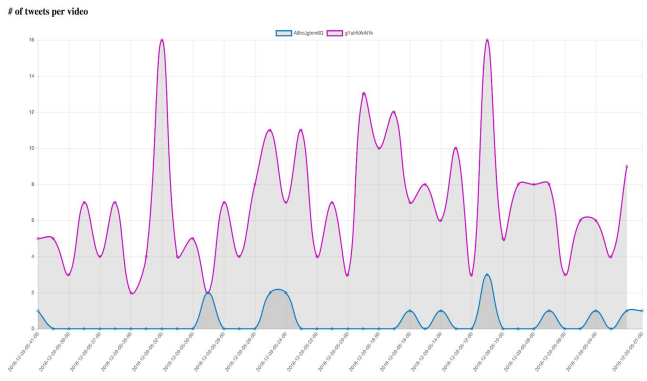

Figure 4.4.1 Real Time Twitter data of trending YouTube videos

From our real time stream we have found that only the most trending videos on YouTube get many tweets about them. Even just one place down from the top they get significantly less tweets. Outside of the top 10 it is unlikely to get any tweets. This may also be due to Twitter limiting the amount of tweets that non paid users can get but we ran this multiple times and acquired the same results. These findings were very unexpected. With how connected the internet and world feel at its current state we expected at least the top 10 videos to have a flurry of tweets about them.

4.5 How trending videos are doing now

Seeing how videos are doing now from when they were trending a year ago met our expectations. The average increase in views since last year was 2328429 and the average increase in likes was 35239 and the average increase in dislikes was 1954. These figures fall in line with our sentiment analysis as most videos are positive therefore there are more likes than dislikes. The disparity between the top and the bottom gainers were enormous. The top 5 gained on average 1.6 billion views since they were on trending, where as the bottom averaged 465 views even though they were on trending.

CHALLENGES AND LIMITATIONS

During the course of this project we ran into a variety of difficulties.

Unfortunately, tracking a video by id on Twitter is not as straightforward as it appears. Since Twitter shortens YouTube URLs into one of three different formats (t.co, youtu.be, or youtube.com) it was difficult to extract the URL from the text. Deeper in the JSON object returned by tweepy, we found an uncompressed URL but it would be indexed into different sections of the JSON depending on the type of tweet. Retweets, Quoted tweets, and even tweets from select users (extended tweets) were all formatted differently. This made it difficult to know where to extract the information without either hard-coding set locations or useing workarounds (such and pinging the shortened URL to get the unexpanded version).

Another limitation we ran into was the limited amount of historical Twitter data we were allowed to access. Since we do not have access to an enterprise level account, the maximum amount of data we could query was 100 tweets from the last 7 days.

Visualization of the streamed data was another point of conflict. Only the top trending videos get a consistent amount of tweets. Less popular videos got little to no tweets or have gaps in between them. Filling in those gaps was difficult because order needed to be maintained.

A major challenge that was faced when performing RGB analysis was how to detect if the image downloaded was a default YouTube thumbnail image. This is difficult to detect since it is no different than any other thumbnail downloaded. To get around this, we automatically filtered our colors that had the exact result of RGB combination of the default image.

FUTURE ANALISIS

Due to the limited time we had for this project, we were unable to pursue some of our more ambitious analytic goals. In the future, if we are to ever return to this project, there are a couple of interesting questions and analysis we would like to delve into:

Object detection for video thumbnails were a failure and is a future pursuit we would like the pursue. Although these

thumbnails where too low resolution to work with. It would be interesting to see what makes a successful thumbnail from colour to content.

Graph likes/dislikes of the time a video is on the trending page. Does a video grow stale after being up on trending for so long?

Mapping out more regions on the world map, we were limited to only 5. It would be interesting to compare vastly different cultures and or get more specific regions like per state instead of United States as a whole.

Since our data spans over 6 months it is unfair to compare videos that had a year to accumulate views vs those who only had 6 months. We took a subset of our data those that were published in the months of November and December of 2017. Using the YouTube api we gathered the current information for those videos and did some simple analytics to see averages of how they did.

ACKNOWLEDGMENTS

Except for the stream data we collect from Twitter and YouTube, all other datasets have been collected and curated by the open source data community at Kaggle. Although we will be performing analysis, we do not own the data. All data used is either collected from YouTube, the Twitter API, or the Kaggle Dataset. The charts used were using source code from chart.js.org and amcharts.com.