Analysis of Trending YouTube Videos' Statistics of United States

A.A.H.W.S. Herath
Department of Computer Science & Engineering
Faculty of Engineering
University of Moratuwa
Email: wikasitha.20@cse.mrt.ac.lk

M.J.I.Ahamed
Department of Computer Science & Engineering
Faculty of Engineering
University of Moratuwa
Email: jiffry.20@cse.mrt.ac.lk

***Abstract*--- YouTube has become a prominent media that has a lot of video content in different categories. Currently it pulls in over 1.8 billion users every day while giving out thousands of trending videos in different countries. YouTube acts as a promising source of income to a lot of its users due to their YouTube channels. Since it has become a highly influencing social media, people are using this platform as a powerful tool to share their thoughts and promote powerful messages and influence globally. In this report we are analyze the US trending videos' dataset in order to find correlations between trending videos' properties. By doing this we are able to come up with some common features in all YouTube trending videos in the US. Furthermore, using this output information from our analysis we can give meaningful suggestions to the people who upload videos on YouTube. We believe by applying our findings, a YouTuber can increase the view count of the video.**

Keywords. You Tube, Analysis, Correlation

1. INTRODUCTION

YouTube was launched in 2005 with the purpose of help people and share videos with a global audience [1]. It was founded by Chad Hurley, Steve Chen, and Jawed Karim, who ran the company from a (de rigueur) small office above a small restaurant in a small Californian city. Afterwards it became as the world's most popular online video site, and almost 5 billion videos are watched now on YouTube every single day. Via YouTube platform, people started to create a video-sharing website on which users could upload, share, and view videos. As a result of this sharing video content has become a cultural phenomenon in current world between kids, teenagers and so and so. And latest statistics show that the traffic to or from YouTube accounts for over 20% of the total web traffic and 10% of the whole Internet traffic [2]. As reported by Alexa, the web traffic monitoring service owned by Amazon, YouTube is the second most popular website globally with over 300 hours of videos uploaded every minute and 5 billion videos watched every single day [3]. Having started in 2005, YouTube has well developed into a leading online video-sharing destination. The millions of video clips on YouTube represent a wide range spectrum of user interests including those of educators, scholars and researchers.

Each You Tube video is belong many specific attributes like title, publish date, channel name, tags etc. and each video is categorized in a specific category which is make easy to find and watch for audience of their desires and aspects. In 2009 there was a research which was conducted to analyze the online video viewership of the US Internet users. That study could find that 38% watches educational videos, 50% of adults in the US tend to watch funny videos, 32% watch TV shows or movies, and 20% watch political videos [7]. In one hand YouTube is one of perfect website for audience to do everything as they wish. In the other hand it is one of business paradise for those who have different kind of skills and seeking to do new things to world. At present most people able to use You Tube as their main earning path or business and even wherever you are in the world anybody is able to create your own YouTube channel and can become a content creator. It is one of outstanding advantage offered to skilled persons by You Tube. In addition to the ease of uploading nearly any kind of video content, viewers or the audience are able to interact with the video content by liking or disliking a video, commenting on a video, commenting on a comment, liking or disliking a comment, or posting a video response Comments on the videos can be used to understand audience's reactions to important issues or toward videos. Audience's feedback is called comments and can be used to mine implicit knowledge about viewers, regions, videos' content, categories, and community interests .As well as You Tube has enabled feature called hash tags and it is one of quick method that can reach out relevant views fast.it makes a video more popular in between worldwide viewers and finally they become as trending videos while making more views and watch hours.

I. Motivation

In this paper we have focused on figuring out statistics of YouTube trending videos in the US. Additionally, we have created a data analyzing process to identify how to make a video as a trending YouTube video and how a trending video depends on likes, dislikes, comments and user views. The data source USvideos.csv data set has been taken from Kaggle and it consists of 40950 rows and 16 columns regarding YouTube videos. Let see our data set and Video id, Trending date, title, channel, category, publish date, tags, views, likes, dislikes, views, comments, thumbnail link, comments disabled, ratings disabled, video error or removed and description are the column names in our data set. At the initial stage it consists of duplicate values, null values and the need to clean before the start of the analyzing process. For better analyzing it requires each video category id and its relevant name.

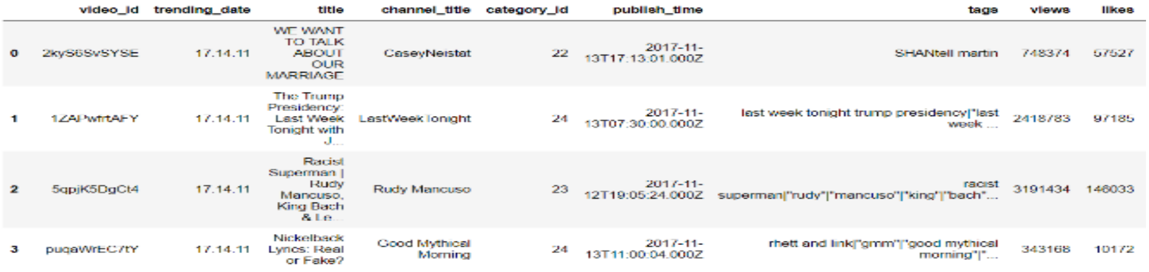
During the analysis we identified correlations between views, likes, dislikes and comments. Later it has been well illustrated on how the relationships behave between the above-mentioned attributes for each YouTube video. After analyzing results, we have produced diagrams and graphs to emphasize our goal and a final decision and predictions. This research will help convey the idea on how to make your own YouTube video as one of the trending Videos in YouTube for those who are seeking to be a youtuber or for those who are interested in YouTube. The results which have been generated by us via this analyze can be used for business planning and strategy by people who do focus on YouTube as a high commercial platform.

II. Data processing

We selected a well-structured dataset for this research from Keggle. Our dataset has been titled as Trending YouTube Videos. There are hundreds of videos which were trending during that particular day and also, they have separated these datasets with different regions for analysis purpose. We have selected the US trending YouTube videos dataset so that we can focus on a single region. When we look at the dataset, unfortunately we only have data from 2017 November to 2018 June. Apart from this, the dataset contains 40950 rows and 16 columns. Columns names are video_id, trendin_date, title, channel_title, category_id, publish_time, tags, views, likes, dislikes, comments count, thumbnail link, comments disabled, ratings disabled, video error_or_removed, descriptions. Each column contains different types of data, some of the columns' data are used for our analysis purposes and some of them are not.

"Fig.1" This is our dataset. In this dataset, we have identified some issues here we are listing those issues

- Trending_date column contains a date which not well-formatted.
- Category_id column contains different category id but not the category name.



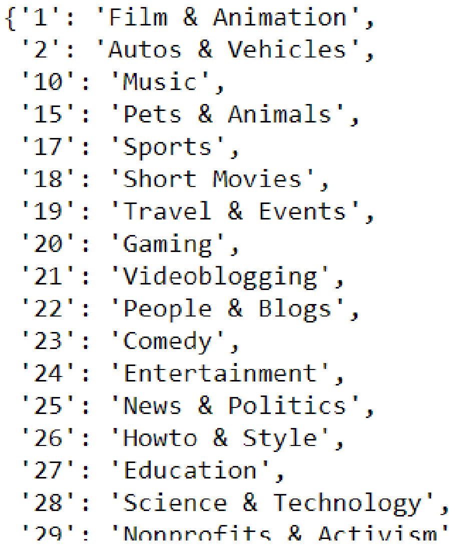
	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes
0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-11-13T17:13:01.000Z	SHANtell martin	748374	57527
1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-13T07:30:00.000Z	last week tonight trump presidency|"last week ...	2418783	97185
2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"king"|"bach"...	3191434	146033
3	puqaWrEC7tY	17.14.11	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	2017-11-13T11:00:04.000Z	rhett and link|"gmm"|"good mythical morning"|"...	343168	10172

Fig. 1 data table _before preprocessing

- We don't need some of these columns for our analysis.
- We need some extra columns of data for further analysis.

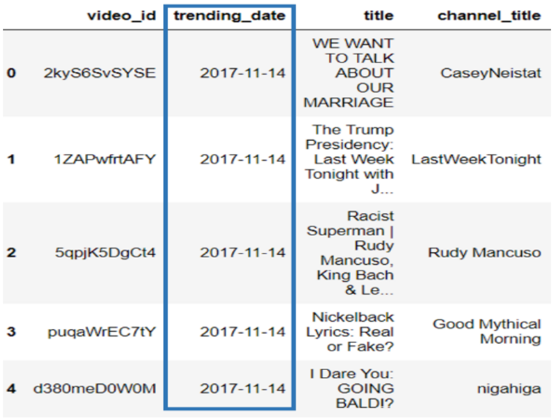
We cleaned our dataset in order to overcome these issues that we have mentioned above.

- "Fig.2" We have formatted our trending date column as a pandas DateTime format.
  - For this, we have used pandas.to_datetime().
- "Fig.3" Category_id column only contains videos' category ids. We need category name only for analysis purpose and we have to map that name with this id all over the dataset. For this purpose, we have created a new CSV file which contains the entire category. After that we have mapped that file with our dataset.
- "Fig.4" shows our category id and name dataset after we mapped with our dataset.
- "Fig.5" Finally, we have removed unwanted columns which are not needed for our analysis purpose. The list of columns we have removed is thumbnail_link, comments disabled, ratings disabled, video error_or_removed, description. However, we did not remove video_id, title, channel_title, tags columns for our future analysis purpose.



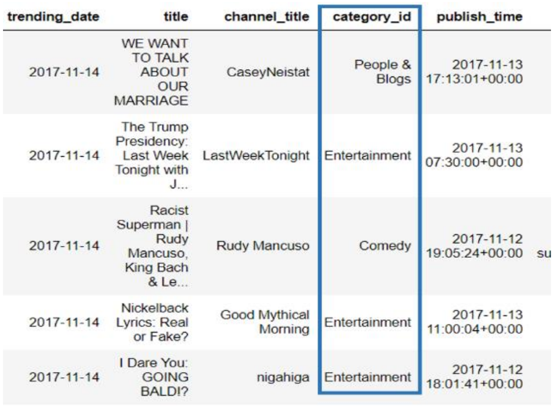
```
{'1': 'Film & Animation',
 '2': 'Autos & Vehicles',
 '10': 'Music',
 '15': 'Pets & Animals',
 '17': 'Sports',
 '18': 'Short Movies',
 '19': 'Travel & Events',
 '20': 'Gaming',
 '21': 'Videoblogging',
 '22': 'People & Blogs',
 '23': 'Comedy',
 '24': 'Entertainment',
 '25': 'News & Politics',
 '26': 'Howto & Style',
 '27': 'Education',
 '28': 'Science & Technology',
 '29': 'Nonprofits & Activism'
```

Fig.3 formatted date column



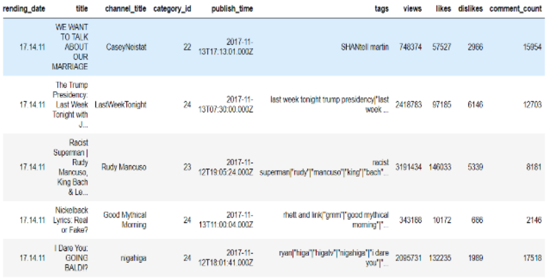
	video_id	trending_date	title	channel_title
0	2kyS6SvSYSE	2017-11-14	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat
1	1ZAPwfrtAFY	2017-11-14	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight
2	5qpjK5DgCt4	2017-11-14	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso
3	puqaWrEC7tY	2017-11-14	Nickelback Lyrics: Real or Fake?	Good Mythical Morning
4	d380meD0W0M	2017-11-14	I Dare You: GOING BALD!?	nigahiga

Fig.2 formatted date column



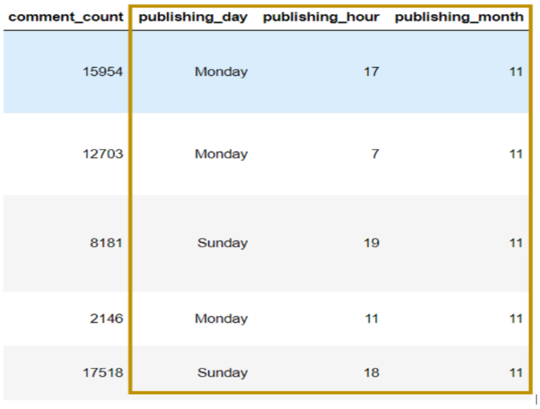
trending_date	title	channel_title	category_id	publish_time
2017-11-14	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	People & Blogs	2017-11-13 17:13:01+00:00
2017-11-14	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	Entertainment	2017-11-13 07:30:00+00:00
2017-11-14	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	Comedy	2017-11-12 19:05:24+00:00
2017-11-14	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	Entertainment	2017-11-13 11:00:04+00:00
2017-11-14	I Dare You: GOING BALD!?	nigahiga	Entertainment	2017-11-12 18:01:41+00:00

Fig.4 Removed unwanted columns



rending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes	comment_count
17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-11-13T17:13:01.000Z	SHANtell martin	748374	57527	2966	15954
17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-13T07:30:00.000Z	last week tonight trump presidency|"last week ...	2418783	97185	6146	12703
17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"king"|"bach"...	3191434	146033	5339	8181
17.14.11	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	2017-11-13T11:00:04.000Z	rhett and link|"gmm"|"good mythical morning"|"...	343168	10172	666	2146
17.14.11	I Dare You: GOING BALD!?	nigahiga	24	2017-11-12T18:01:41.000Z	ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...	2095731	132235	1989	17518

Fig.5 Category name



comment_count	publishing_day	publishing_hour	publishing_month
15954	Monday	17	11
12703	Monday	7	11
8181	Sunday	19	11
2146	Monday	11	11
17518	Sunday	18	11

Fig.6 Publish_date_hour separately

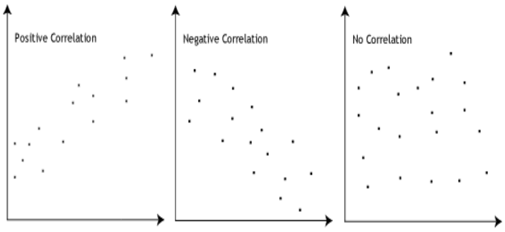
- “Fig.6” Apart from these columns which already came up with dataset we have created some new columns with other existing data for our further analysis purpose, these columns have been created whenever they needed for the analysis purpose. These are columns we have created publishing day, publishing hour and pulishing_month.

III. Methodology

In this paper, we are conducting a study on trending YouTube videos’ properties in order to find some meaningful insights. We have used the Pearson Correlation Coefficient to find the correlations between videos’ views, like, dislike and comments_cout.

“Fig.8” What is the Pearson Correlation Coefficient?

Pearson Correlation coefficient is used to find the linear correlations or association between two variables. It can take a range of values from +1 to -1 and if it’s 0 then there is no relationship between them. A value greater than 0 indicates a positive association and a value less than 0. indicates a negative association. It indicates the strength of the relationship.



. Fig.7 Correlation types

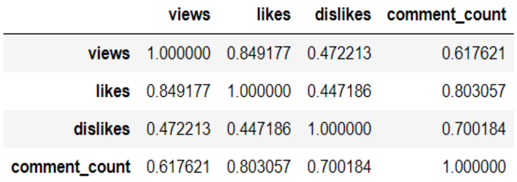
In this paper, we find the Pearson correlations coefficient between these views, like, dislike and comments_cout property of training videos so that we can identify an impact on one variable in another using the formula below.

Apart from this correlation coefficient analysis, we can have some different analysis as well to find more insights for the following purposes.

- trending videos’ category with its views and likes
- Which day of the week most of the trending videos have been uploaded
- Which time in a day most of the trending videos have been uploaded

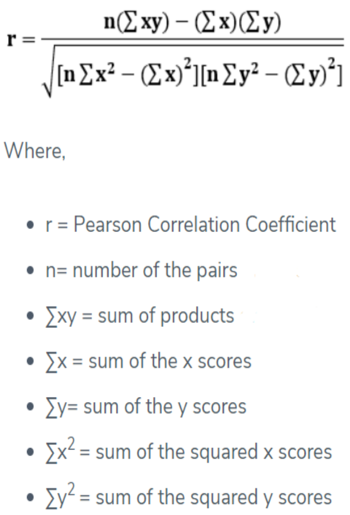
IV. Results and Conclusion

“Fig.9” We found the Pearson correlations coefficient between columns views, like, dislike and comments_cout. From the above table, views-likes shows the highest Pearson correlations coefficient value, after this views-comments_count shows the second-highest value, but views-dislikes shows the least value. Therefore, it can be concluded that the likes column has the most impact on views column apart from the other two properties of the video This result gives a clear view that if any video becomes trending it mostly depends on how people like that video.



	views	likes	dislikes	comment_count
views	1.000000	0.849177	0.472213	0.617621
likes	0.849177	1.000000	0.447186	0.803057
dislikes	0.472213	0.447186	1.000000	0.700184
comment_count	0.617621	0.803057	0.700184	1.000000

Fig.9 pearson correlation coefficient



$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where,

- r = Pearson Correlation Coefficient
- n= number of the pairs
- ∑xy = sum of products
- ∑x = sum of the x scores
- ∑y= sum of the y scores
- $\sum x^2$ = sum of the squared x scores
- $\sum y^2$ = sum of the squared y scores

Fig.8 Pearson correlation equation

“Fig.10” If any video satisfies the audience, then it will become a trending one. Apart from this finding, one more conclusion can be derived. If we look at comments_count-likes and comments_count-dislike, we can identify both have a nearly same value. It indicates that even if people like it or not they tend to give their opinion. Apart from the above table, we can come up with the below pair plot

"Fig.11" We have analyzed the video category with views and likes so that we can get some more details. With details on which category video gets more views and likes we plotted these details in the below. we can see that most people viewed and liked entertainment videos. To compare between different video category, we came up with this below chart.

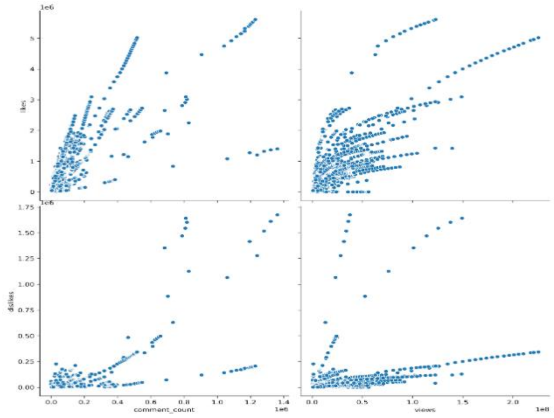


Fig.10 Pairplot

"Fig.12" we are able to see that the Entertainment category has the highest number of trending videos. Here we have noticed one more aspect which is that entertainment is more than four times of education videos. Furthermore, we came up with some different analysis parts. We have found each video published in days of the week and gotten the total count for each day and plotted it on a bar chart as shown in the figure below.

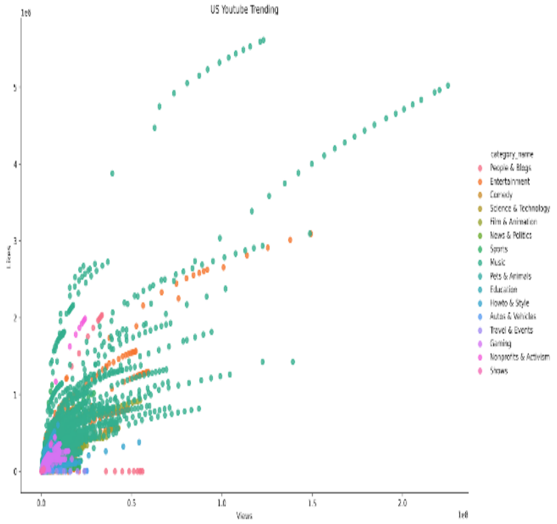


Fig.11 Video category with its likes and views

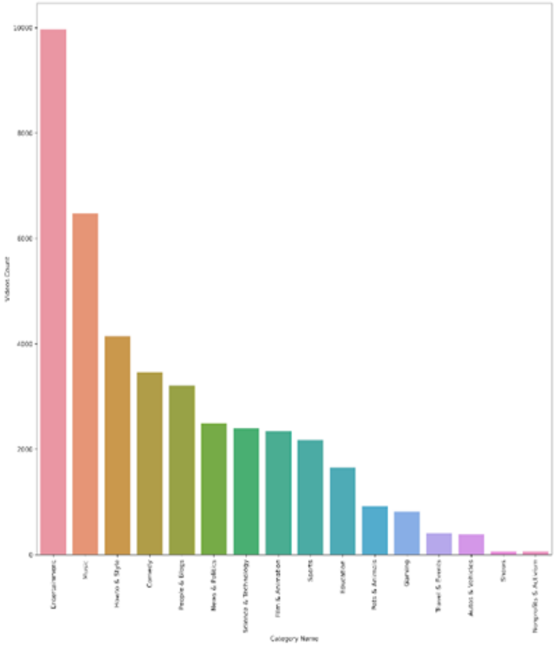


Fig.12 Video category comparison

"Fig.13" We can see that most of the trending videos have been uploaded on Friday and the least amount of videos have been uploaded on Saturday. There might be some different reasons for this insight. On Friday's people get off from work and have more free time during weekends. As a result, they watch and spend more time on YouTube. After this, we came up with another analysis. We found the total count of publishing hours and we plotted it into a bar chart.

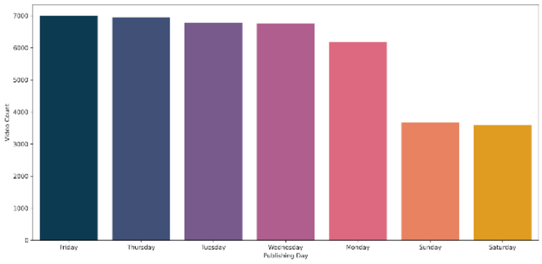


Fig.13 Trending videos uploading times

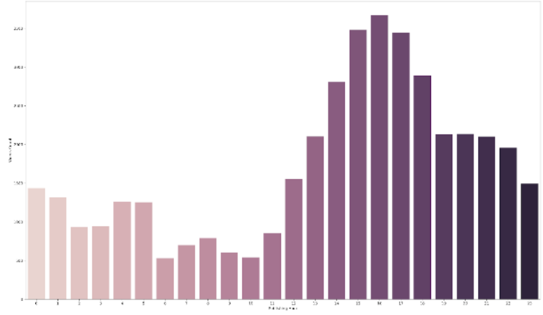


Fig.14 Total count of publishing hours

Fig.14 We can clearly identify most of the trending videos have been uploaded during 3 PM to 5 PM. All these above analysis and findings tell a lot of information about the US YouTube trending videos. After all these findings we can arrive at a good decision on how to make a video as a trending video. Since YouTube became a good source of income and influence, media people are making tons of videos every single day. We can help a YouTuber to make their video a trending one by telling them what day, time and category they have to upload it on. Apart from this, we can say a video becomes a trending one when people like that video as it is all about audience satisfaction.

V. ACKNOWLEDGEMENT

We would like to express our gratitude to our primary supervisors Dr. Shehan Perera and to Dr. Uthayasanker Thayasivam for their support and guidance to carry out this analytical study. We would also like to thank our friends and family who supported us and offered deep insight into the study.

VI. FUTURE WORK

In future, we can conduct and analyze this topic in depth using tags, descriptions and thumbnail columns related to YouTube. It will be useful to find the most used tags on trending videos for each video category. Similarly, we can conduct further analysis on descriptions and thumbnails as well.

.

REFERENCES

[1] T. Anderson, & H. Kanuka, "E-research: methods," strategies and issues. New York: Allyn & Bacon, 2003.

[2] A. Lenhart , K. Purcell, A. Smith, "Social Media & Mobile Internet Use Among Teens and Young Adults," *pew Internet and American Life Project*, 2010.

[3] C. Snelson, "Web-based video for e-learning: Tapping into the YouTube phenomenon.," *Implications of web-based communities and networking*, pp. pp.147-166), 2009.

[4] C. Snelson, "YouTube across the disciplines: A review of the literature," 2010.

[5] D. Derks, A. Bos, V. Grumbkow, J., "Emoticons and online message interpretation," *Social Science Computer Review*, vol. 26(3), pp. 379-388, 2008.

[6] H. Mike, S. Pardeep & Vis, Farida., "Commenting on YouTube Videos: From Guatemalan Rock to El Big Bang," *From Guatemalan Rock to El Big Bang. Journal of the American Society for Information Science and Technology.*, vol. 63, no. 10.1002/asi.21679, pp. 616-629, 2010.

[7] J. Hopkins, "Surprise! There's a third YouTube co-founder," *USA Today*, no. July 15, 2011.

[8] K. Karsten, Z. Carmen, & W. Friedrich, "Leveraging the affordance of YouTube: The role of pedagogical knowledge and mental models of technology functions for lesson planning with technology.," *Computers & Education*, vol. 58, pp. 1194-1206, 2012.

[9] K. Thorson, B. Ekdale, P. Borah, K. Namkoong, "YouTube and Proposition 8: A case study in video activism.," *Information, Communication & Society*, pp. 325-349, 2010.

[10] P. Steinberg, L.Wason, S. Stern, J. M., Deters, "YouTube as source of prostate cancer information," *Urology*, vol. 75(3), pp. 619-622, 2010.

[11] R. Chenail., "YouTube as a qualitative research asset: reviewing user generated videos as learning resources," *The Weekly Qualitative Reports*, vol. 1(4), pp. 18-24.

[12] R. Ibrahim, B. Yusliza, "Jangkaan masa depan homeschooling di Malaysia dan impaknya terhadap kurikulum di Malaysia," *International Conference on World Class Education*, 2011.

[13] R. Maznah, R. Hussain, *Empowering learners as the owners of feedback while YouTube-ing Interactive Technology & Smart Education Journal*, vol. 6(4), pp. 274-285.

[14] R. Mullen, & L. Wedwick,"Avoiding the digital abyss," *Getting started in the classroom with youtube, digital stories, and blogs. Clearing House,*, vol. 82(2), pp. 66-9, 2008.