Minor Project Report
on

Data analysis of trending YouTube videos

Submitted to
Amity School of Engineering and Technology
By

Tarun Nalwa

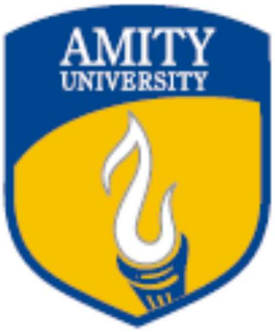
(A2305215104)

Shubham Kathuria

(A2305215527)

Under the guidance of:

Dr. Praveen Kumar



AMITY UNIVERSITY UTTAR PRADESH
AMITY SCHOOL OF ENGINEERING AND TECHNOLOGY
SECTOR-125, NOIDA- 2013

DECLARATION

I **Tarun Nalwa** and **Shubham kathuria**, student(s) of B.Tech (CSE) hereby declare that the project titled “Data Ananlysis of Trending Youtube Videos ” which is submitted by me/us to Department of B.Tech(CSE), **Amity School of Engineering and Technology (ASET)** , Amity University Uttar Pradesh, Noida, in partial fulfillment of requirement for the award of the degree of Bachelor of Technology in CSE , has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition.

The Author attests that permission has been obtained for the use of any copy righted material appearing in the Dissertation / Project report other than brief excerpts requiring only proper acknowledgement in scholarly writing and all such use is acknowledged.

Noida
Date: 12/19/2018 **Name and Signature:**

CERTIFICATE

On the basis of declaration submitted by **Tarun Nalwa** and **Shubham Kathuria**, student(s) of B. Tech CSE, I hereby certify that the project titled "Data Ananlysis of Trending Youtube Videos" which is submitted to Department of Computer Science and Technology, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in CSE, is an original contribution with existing knowledge and faithful record of work carried out by him/them under my guidance and supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Date: 12/19/2018 **Dr. Praveen Kumar**

Noida

ACKNOWLEDGEMENT

The key elements concentration, dedication, hard work and application are not only essential factors for achieving the desired goals but also guidance, assistance and co-operation of people is necessary.

I am grateful to my faculty guide **Dr. Praveen Kumar,** Assistant Professor, Amity School of Engineering and Technology (ASET) for guiding me. He has the qualities to be an ideal guide. He will always prevail upon my remembrance.

I am also grateful to our respectful head of department **Prof. (Dr.) Abhay Bansal** department of Computer Science & Engineering for their regular encouragement and endeavors.

I am very fortunate to have unconditional support from my family. I thank my parents, who gave me the courage to get my education, supported me in all achievements throughout my life. Without their encouragement, this work would indeed have been very difficult for me to tackle. Above all, I pay my reverence to the almighty GOD.

TARUN NALWA
(A2305215104)
Shubham Kathuria
(A2305215527)

ABSTRACT

YouTube is a compendious video source on the internet and the videos on YouTube have different statistics compared to traditional streaming videos, in terms of length, access pattern, etc. Moreover, the developers can access video statistics through its Application Programming Interface. With the advent of YouTube API (v3) functionality of YouTube can be even incorporated in our own applications. The attributes that make a YouTube video popular can be determined by studying and exploring the datasets of trending YouTube videos. In this project, we investigate the impact of : views, likes and dislikes, comment_count, dislike_percentage, days_to_trending and video_category. The data sets of those videos were taken from YouTube API (v3). The dataset contains only metadata and no data like video, image, audio, or large text documents. We analyze these datasets to determine the impact of likes, dislikes, views and number of comments on the trending date of a YouTube video. We also explore the relationship between user comment behavior and how it might or might not be predictive of video consumption patterns. Keywords: Data Analysis, YouTube API, JSON, R, Prediction, Sentiment Analysis, Python.

TABLE OF CONTENTS

INTRODUCTION

YouTube is a platform which allows discovering, viewing, commenting, giving a feedback and sharing videos. It was founded in February 2005 and according to the statistics of youtube, more than 1.5 billion unique people visited youtube in May 2018. YouTube is localized in more than 91 countries and covers 80 languages. The number of daily subscriptions have been rising exponentially since a few years. The number of users that watch branded videos on the internet is quite high and the usage of Youtube for this purpose is comparable to Facebook and Instagram. YouTube maintains a list of the top trending videos on the platform. To determine the top-trending videos, a combination of factors including measuring users interactions like number of views, shares, comments and likes is used. Top performers on the YouTube trending list are music videos celebrity and/or reality TV performances, and the random dude-with-a-camera viral videos that YouTube is well-known for. Videos can be uploaded on YouTube in .WMV, .AVI, .MOV and .MP formats but YouTube converts them into .FLV (Adobe Flash Video)format after uploading.

The steps involved in studying and exploring the datasets of trending YouTube videos are as follows:

1. Obtaining Google developer API key
2. Collecting data using YouTube video IDs
3. Saving and reading YouTube data file

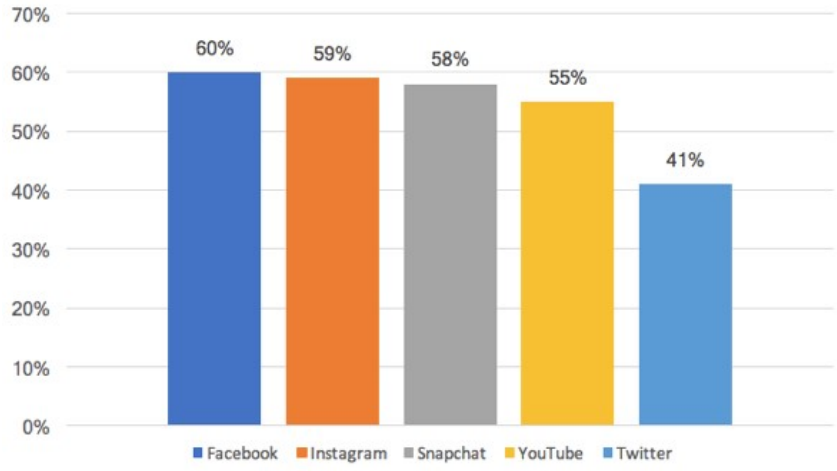


Fig.1. Consumers watching branded videos daily

4. Creating user network
5. Histogram of node degree
6. Obtaining sentiment scores of comments on YouTube videos by users

The dataset that we use contains data of several months of trending youtube videos. It contains data of India, USA, Canada and Great Britain. Each country's data is stored in different files and the data includes title of the video and its channel, likes, dislikes, description, count of comments, publish time, tags, trending date. The data also includes a category id field, which is unique for each region. The categories for a specific video can be retrieved from the associated JSON.

LITERATURE REVIEW

YouTube is one of the world's biggest video sharing stages, where recordings are transferring persistently by the a great many clients (organizations, private people, and so on.) [18]. YouTube has developed as a complete and open accumulation of video data source on the web. It is an exceptional situation with numerous aspects, for example, multi-modular, multi-lingual, multidomain and multi-social [18]

YouTube was positioned as the second most mainstream website by Alexa Internet, a web movement examination organization on Dec, 2016 [19]. With the end goal to build the client's communication it enables their clients to express their conclusion by rating the seen items (by means of tapping on the like/dislike catches) and cooperating with the other network individuals (through the remarks include) [20].

These exercises (like/despise/number of perspectives) of the clients can fill in as a worldwide marker of value or prominence for a specific video [20, 21]. Besides, these Meta information (like/dislike/number of comments) effectively help the network to channel applicable conclusions all the more productively [20, 22, 23].

With the end goal to locate the ideal and significant video, a compelling and effective process appears to be possible which would not rely upon just those metadata (like/dislike/number of comments). With the end goal to unravel these issues a few works have done where a few gatherings have been accounted for noteworthy advancement [20, 21, 24, 25].

Estimation investigation is the field that reviews and breaks down individuals' reactions and acknowledgment towards a substance (e.g. web journals, items, books, recordings … .) utilizing content investigation computational and calculations to help decide individuals literary responses in the event that they are certain, negative or impartial [1]. Utilizing estimation examination, changes in stock costs can be anticipated [2], political race inclinations can be watched [3], and even radicalization gatherings' connections can be followed among numerous other immediate and circuitous advantages [4,5].

People and associations discover these procedures asset and tedious. These procedures can even outcome in one-sided assessments, since they are normally assembled from a little gathering of buyers, or from the ones who will give their suppositions [6]. New patterns have showed up as of

late in the slant investigation inquire about field where the examination of creators suppositions isn't just done literarily, yet in addition is done auditorily and outwardly [7, 8].

Numerous explores constructed their work with respect to extricated datasets from Twitter, and examined them utilizing propelled machine dialect calculations like Support Vector Machines (SVM) [9, 10], Naïve Bayes (NB), and K-Nearest Neighbor (KNN) [11, 12].Most of the English dialect opinion investigation analysts led their feeling examinations on their corpuses utilizing the SentiWordnet dictionary [13] . [14] exhibited physically explained English and Italian corpuses, and an investigation of the significance between the video's remarks and the video substance. To bind their methodology, they gathered their video sets focusing on extraordinary brands of autos, tablets and advanced cameras

The writers in [15] gathered 4050 English surveys/remarks about Academic, News and business points from Twitter, Facebook and YouTube with the end goal to gauge the precision of the English opinion dictionaries by applying K-Nearest Neighbor (KNN), Naïve Bayes (NB) and Support Vector Machines (SVM) grouping strategies.

Writing survey uncovers that estimation investigation has been normally been carried on printed information, for example, in [16].carrying out a comparable errand on video presents high volumes of information taking care of and extricating important data from the video content, which is a non-unimportant issue. This region of research is picking up significance because of the headways and accessibility of computerization devices [17].

The exploration network ceaselessly indicated distinct fascination in examining and abusing the rich substance shared on YouTube. Some prior examinations endeavored to explore the recovery capability of the video utilizing not just the Meta information (like/dislikes/number of comments) [18] yet additionally utilizing remarks [20, 24].

METHODOLOGY

In order to store, study and analyse data of trending youtube videos, we need to store in tabular form. The data can be saved in excel or csv. Unlike excel sheets, csv is a plain text in which the values are separated by commas. We use csv format files in this project. The data of trending videos collected from youtube is distinguished in the following: Comment, User, ReplyCount, LikeCount, PublishTime, CommentId, ParentID, ReplyToAnotherUser, VideoID.

Meta data of a youtube video includes the following :

- ID
- Uploader
- Date Added
- Category
- Video Length (in seconds)
- Number of Views
- Number of Comments
- Related Videos

Technologies Used:

- Python
- JSON
- R

Python is one of the easiest programming languages and is widely used nowadays as we can do almost anything with it. Certain libraraies such as numpy, pandas, matplotlib.pyplot, seaborn have been used in the project to get the data of trending youtube videos.

JSON or JavaScript Object Notation is a lightweight open standard data interchange file format. It is a subset of JavaScript Programming Language and it is used to transmit data between a web application and a server. It works in a similar way as XML.

R is a free software environment that is used for statistical computing and graphics and works on Windows, Linux and Mac-OS. It is widely used for analytics, data mining, and data science. RStudio is a user friendly environment for R that has become popular. We used R for calculating the sentiment scores. We used vosonSML package to retrieve data from YouTube and further calculated the sentiment scores of YouTube videos.

PROPOSED WORK

Using Python and JSON

To get youtube data we need to have a Google Developer API Key. The key is needed for obtaining credentials of authorization. It can be generated through google developers console. API key is unique for each user. To get the api key first we need to create a project in :

https://console.developers.google.com

Then the api key can be further copied from the credentials tab. Data of youtube videos can further be retrieved using the following URL :

https://www.googleapis.com/youtube/v3/activities?part=snippet,contentDetails&channelId=<Your-Channel_ID>&key=<Your_API_Key>

Getting Data of comments of a particular trending youtube video :

To get data of comments on any particular trending video we can use edit the following URL by inserting Video_ID of that video and our API Key :

https://www.youtube.com/redirect?q=https%3A%2F%2Fwww.googleapis.com%2Fyoutube%2Fv3%2FcommentThreads%3Fpart%3Dsnippet%26videoId%3DVideo_id%26key%3DYOUR_API_KEY&v=AcUauzCn7RE&event=video_description&redir_token=fnr7fae4X7plzz6OBKmf-3qn_1l8MTU0MjAxNjYwM0AxNTQxOTMwMjAz

Using the above link we retrieved the data of a trending youtube video in India with title: “Top 10 Unlucky Run Outs in Cricket History - Unluckiest Run Outs”. Data of one of the comments on this video is as follows:


```
{
   "kind": "youtube#commentThread",
   "etag": "\"XI7nbFXulYBIpL0ayR_gDh3eu1k/JYtVzYBs9lvbtddn1RmgBzqIgeQ\"",
   "id": "UgwGycQ5XphV1vnIMuh4AaABAg",
   "snippet": {
    "videoId": "PynSVIbIpfc",
    "topLevelComment": {
     "kind": "youtube#comment",
     "etag": "\"XI7nbFXulYBIpL0ayR_gDh3eu1k/r2e9p-
kbEWq_rxXyQjw2OJsA_T0\"",
```

```
     "id": "UgwGycQ5XphV1vnIMuh4AaABAg",
     "snippet": {
      "authorDisplayName": "Aanya Sonani",



"authorProfileImageUrl": "https://yt3.ggpht.com/-
JHiintER88A/AAAAAAAAAAI/AAAAAAAAAAA/PPrCBFHxQ-s/s28-c-k-no-mo-rj-
c0xffffff/photo.jpg",
      "authorChannelUrl":
"http://www.youtube.com/channel/UC8ctRpqrQclTl0Tqu1oQE_w",
      "authorChannelId": {
       "value": "UC8ctRpqrQclTl0Tqu1oQE_w"
      },
      "videoId": "PynSVIbIpfc",
      "textDisplay": "I believe that all this is good luck",
      "textOriginal": "I believe that all this is good luck",
      "canRate": true,
      "viewerRating": "none",
      "likeCount": 1,
      "publishedAt": "2018-11-11T09:36:37.000Z",
      "updatedAt": "2018-11-11T09:36:37.000Z"
     }
    },
    "canReply": true,
    "totalReplyCount": 0,
    "isPublic": true
   }
  }
```


Certain libraries of python such as numpy, pandas, matplotlib.pyplot, seaborn can be used to get the data of trending youtube videos. Finding the relation between Category_ID and number of videos uploaded on that category.

```
import numpy
import pandas
import seaborn
seabors.set(font_scale=1.5,rc={'figure.figsize':(11.7,8.27)})

import matplotlib.pyplot
import matplotlib.colors
import datetime
```

```
# Import Data of Trending Videos in India
ind_videos = pandas.read_csv('../input/INDvideos.csv')
ind_videos_categories = pandas.read_json('../input/IND_category_id.json')
```

TABLE.1. Data of Trending Youtube Videos

	views	likes	dislikes	comment_count
count	24951.0	24951.0	24951.0	24951.0
mean	1343968.8	46985.5	2815.3	5954.5
std	4218124.8	150213.7	34503.7	32214.7
min	549.0	0.0	0.0	0.0
5%	16639.0	133.0	9.0	20.0
25%	128213.5	2598.0	115.0	354.5
50%	380393.0	10745.0	379.0	1165.0
75%	1120025.5	32588.5	1233.0	3655.0
95%	4794933.0	184515.0	7241.5	20422.0
max	149376127.0	3093544.0	1674420.0	1361580.0

```
# Map Category IDs using the supporting file: US_category_id.json
categories = {int(category['id']): category['snippet']['title'] for category in in_videos_categories['items']}
ind_videos.describe(percentiles=[.05,.25,.5,.75,.95]).round(1)
```

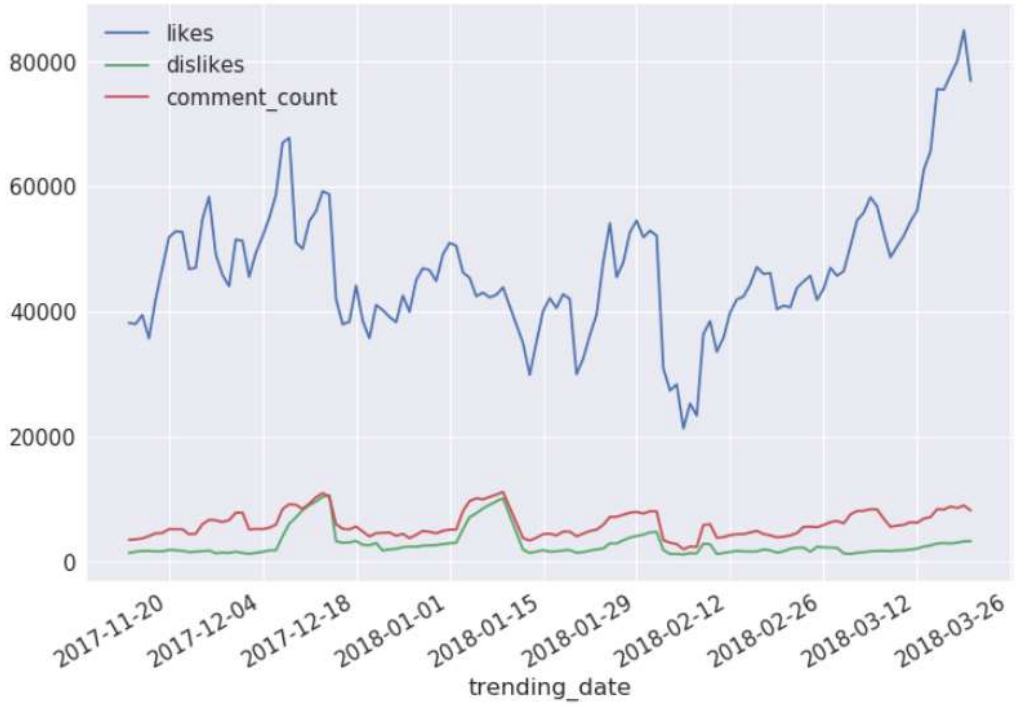


Fig.2. Likes, Dislikes, Comment_Count vs Trending date

```
table = pd.pivot_table(us_videos, index=us_videos.index.labels[0])
table.index = us_videos.index.levels[0]
_ = table[['likes','dislikes','comment_count']].plot()
_ = table[['views']].plot()
```

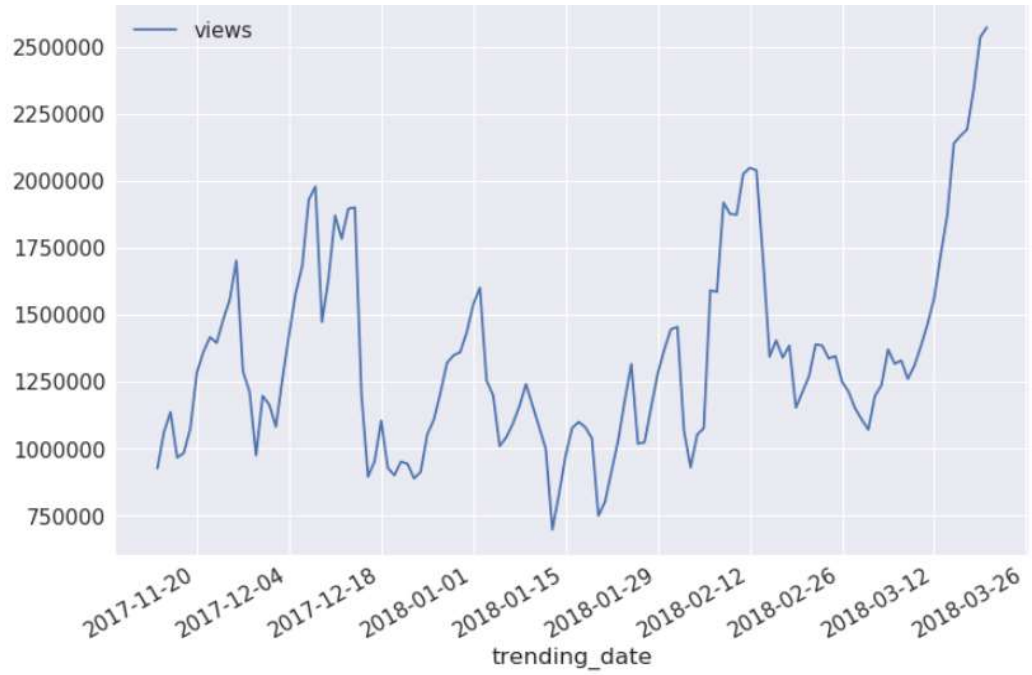


Fig.3. Views on a YouTube Videos vs trending date

The relation between Category_ID and number of videos uploaded on that category can be determined using following code :

```
cat_id_mapping = {2:'Autos & Vehicles',1:'Film & Animation',
                  10:'Music',15:'Pets & Animals',17:'Sports',
                 19:'Travel & Events',20:'Gaming',22:'People & Blogs',
                 23:'Comedy',24:'Entertainment',25:'News & Politics',
                 26:'Howto & Style',27:'Education',28:'Science & Technology',
                  29:'Nonprofits & Activism',43:'Shows'}
df_videos_gb = df_videos.groupby('category_id').count()['title']
df_videos_gb = df_videos_gb.rename(cat_id_mapping)
ax = df_videos_gb.plot(kind='bar',title='Video Categories by their Count',color='green',f
igsize=(10,5))
ax.set_xlabel('Category')
ax.set_ylabel('Count')
```

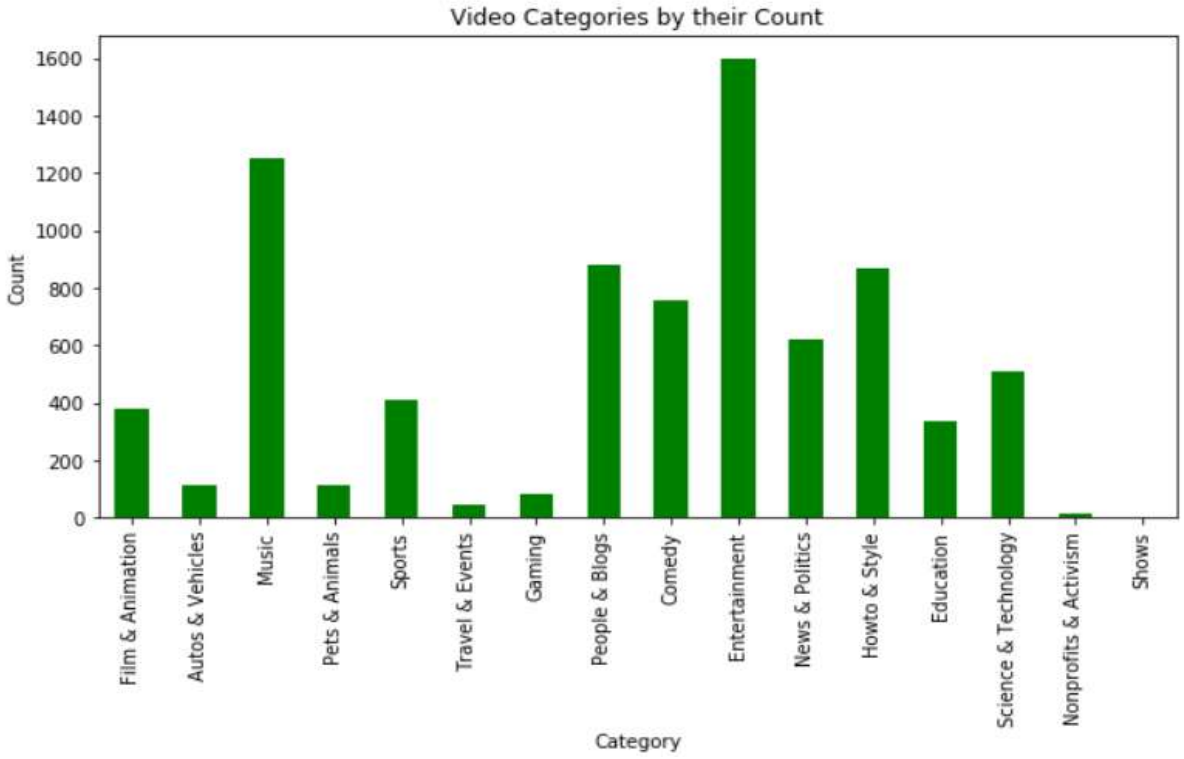


Fig. 4. Relation between Video Category and No. of Videos of that category

The relation between trending video like & dislike and their views & comments can be found out using the following code in python:

```
import numpy
import pandas
import matplotlib.pyplot
import seaborn
from wordcloud import STOPWORDS,WordCloud
sns.pairplot(df_videos,x_vars=['comment_total','views'],y_vars=['likes','dislikes'],size=5)
```

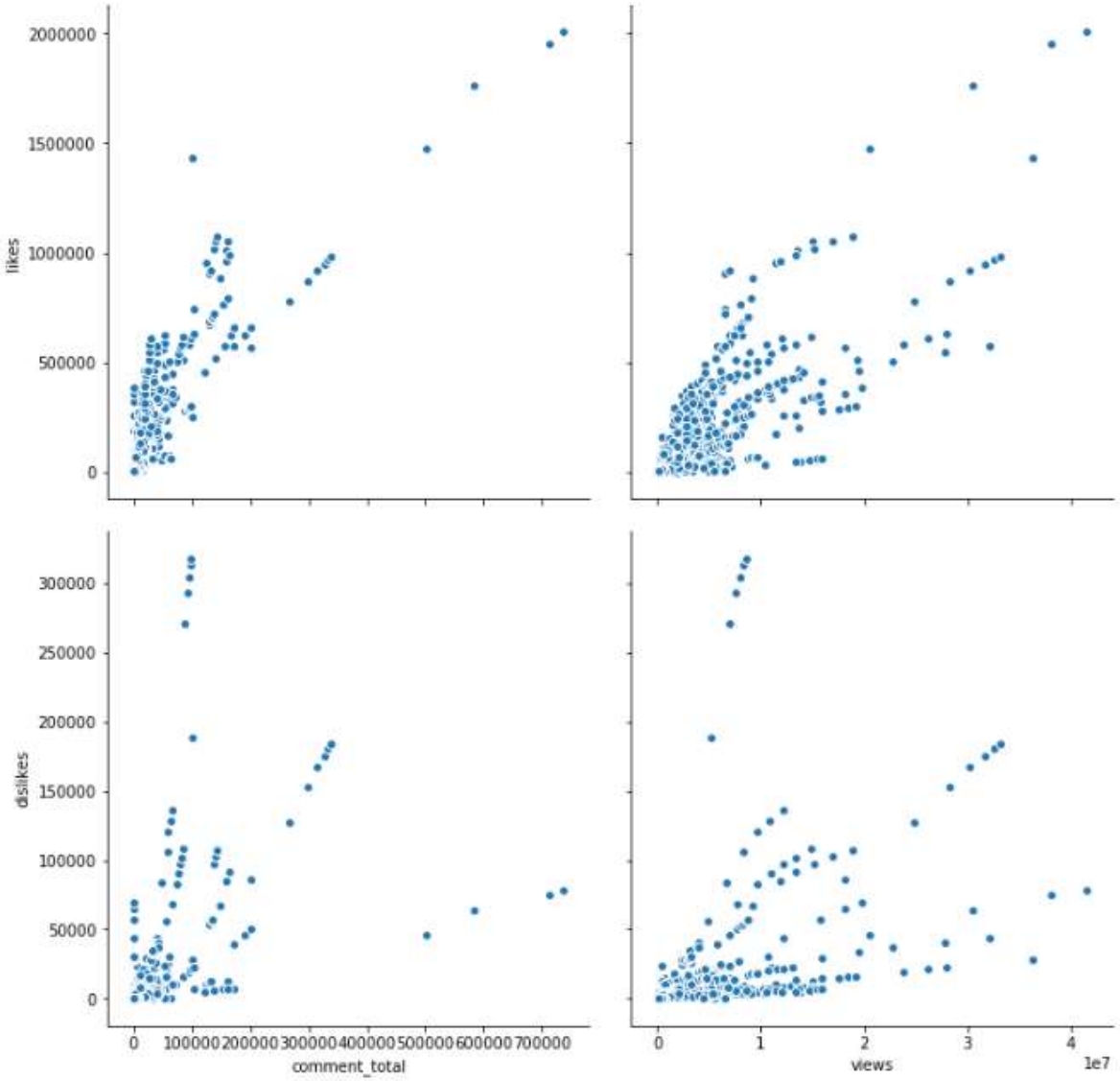


Fig.5. The relation between trending video like & dislike and their views & comments

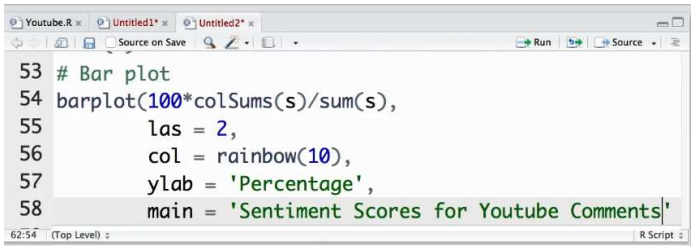
Using R language

vosonSML package of R can also be used to get YouTube data. The data that we read from the csv file is stored by R as dataframe.

```
#Collecting data using Youtube video IDs in R
apikey <- "xxxxxxxx"
AuthenticateWithYoutube(apikey)
```

```
video <- c('5eDqRysaico','dJclNIN-Tpo&t=25s')
ydata <- CollectDataYoutube(video, key, writeToFile= FALSE)
str(ytdata)
write.csv(ydata, file='~/Desktop/yt.csv', row.names=F)

#Read Youtube Data File
data<- read.csv(file.choose(), header = T)
str(data)
```



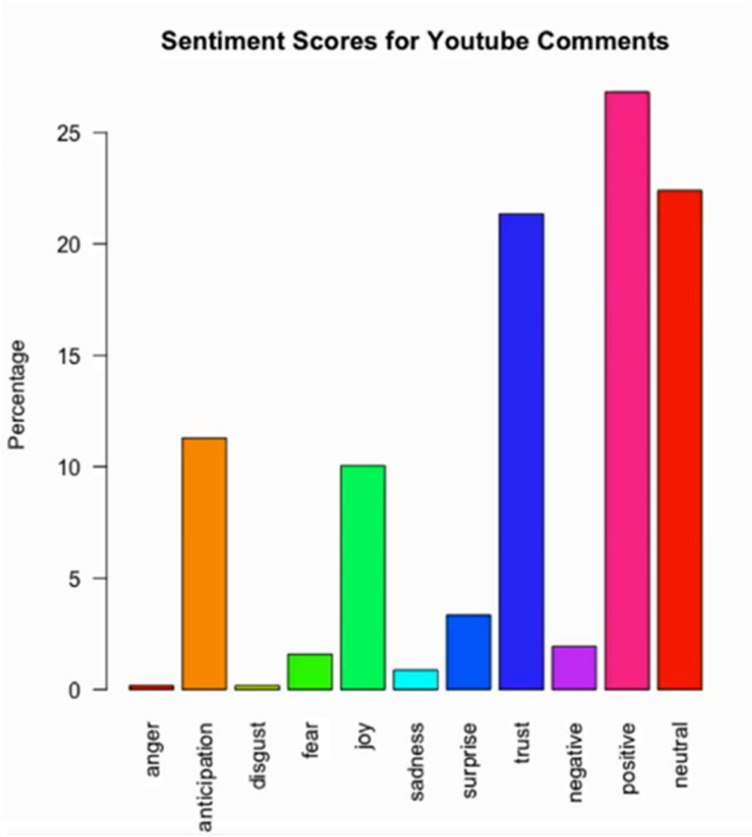


Fig.6. Sentiment Scores for YouTube comments using Data Analytics tools

RESULTS

From the analysis, we find that the popularity of YouTube videos depends on certain factors like category, upload time, likes/dislikes, comments, channel subscriptions, etc. The relation between trending video like & dislike and their views & comments is also shown above in Fig.5. The data sets of those videos were taken from YouTube API (v3) and stored into csv files. Analysing the data sets we find the the relation between Category_ID and number of videos uploaded on that category. We have also shown that how Likes, Dislikes and the number of comments on a YouTube Video make an impact on its trending date. The sentiment scores have been calculated using RStudio by analyzing the data of top trending Youtube videos. We also find from the sentiment analysis of of the user comments on trending YouTube videos that the users found those videos positive, trustworthy and that the videos which have anger disgust and sadness have less chance of making up to top trending list of YouTube. We also get to the conclusion that the time that a YouTube video takes to make it to the trending list is between 0-8 in most countries. But it does vary for some countries.

REFERENCES

[1] B. Pang and L. Lee, "Sentiment Analysis and Opinion Mining" Foundations and Trends in Information Retrieval, 2(1-2), pp. 1135, 2008.

[2] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in the 4th International Conference on Weblogs and Social Media (ICWSM-2010), 2010.

[3] A. Bermingham, M. Conway, L. McInerney, N. O'Hare, and A. Smeaton, "Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation," in the International Conference on Advances in Social Network Analysis and Mining (ASONAM '09), pp. 231-236, 2009.

[4] J. Karlgren, M. Sahlgren, F. Olsson, F. Espinoza, and O. Hamfors"Usefulness of Sentiment Analysis ," in the 34th European conference on Advances in Information Retrieval, 2012.

[5] B. Liu, "Sentiment Analysis and Opinion Mining," Morgan and Claypool, 2012.

[6] H. P. Luhn, "A Business Intelligence System," IBM Journal of Research and Development, 1958.

[7] S. Poria, E. Cambria, N. Howard, G. Huang and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," in Neurocomputing, 174, pp. 50-59, 2016, Elsevier.

[8] M. Wollmer, F. Weninger, T. Knaup, B. Schuller, S. Congkai, K. Sagae and L. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," in IEEE Intell. Syst., vol. 28, no. 3, pp. 46-53, 2013

[9] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in 2012 International Conference on Collaboration Technologies and Systems (CTS), IEEE, 2012.

[10] M. Nabil, M. Aly, and A. Atiya, "LABR: A Large Scale Arabic Sentiment Analysis Benchmark," 2015.

[11] R. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat, "Sentiment Analysis in Arabic tweets," in 5th International Conference on Information and Communication Systems (ICICS), 2014.

[12] R. Duwairi, "Arabic Sentiment Analysis Using Supervised Classification," in the IEEE 2nd International Conference on Future Internet of Things and Cloud (FiCloud-2014), Spain, 2014.

[13] SentiWordNet, [online]. Available: http://sentiwordnet.isti.cnr.it/ . [Accessed: 17- Aug- 2016]
[26] O. Uryupina, B. Plank, A. Moschitti, A. Rotondi, and A. Severyn, "SenTube: A corpus for sentiment analysis on YouTube social media," in LREC, 2014.

[14] O. Uryupina, B. Plank, A. Moschitti, A. Rotondi, and A. Severyn, "SenTube: A corpus for sentiment analysis on YouTube social media," in LREC, 2014

[15] M. Al-Kabi, N. M. Al-Qudah, I. Alsmadi, M. Dabour, and H. Wahsheh, "Arabic / English Sentiment Analysis: An Empirical Study, " in the 4th International Conference on Information and Communication Systems (ICICS 2013), 2013.

[16] Raja Ashok Bolla," Crime pattern detection using online social media", MSc Thesis, Missouri University of Science and Technology, US, 2014

[17] Wootton, Cliff, "Developing quality metadata: building innovative tools and workflow solutions", Focal Press, 2009, ISBN-13:978-0240-80869-7.

[18] A. Severyn, A. Moschitti, O. Uryupina, B. Plank and K. Filippova,"Multi-lingual opinion mining on youtube," Information Processing & Management, 52(1), 2016, pp. 46-60.

[19] S. Chelaru, C. Orellana-Rodriguez and I. S. Altingovde, "How useful is social feedback for learning to rank YouTube videos?" In World Wide Web, 17(5), 2013, pp. 1-29.

[20] P. Schultes, V. Dorner and F. Lehner, "Leave a Comment! An InDepth Analysis of User Comments on YouTube," Wirtschaftsinformatik, 2013, pp. 659-673.

[21] S. Siersdorfer, S. Chelaru, J. S. Pedro, I. S. Altingovde and W. Nejdl, "Analyzing and mining comments and comment ratings on the social web," ACM Transactions on the Web (TWEB), 8(3),

2014, pp. 1-39.

[22] S. Siersdorfer, S. Chelaru, W. Nejdl and J. San Pedro, “How useful are your comments? analyzing and predicting youtube comments and comment ratings,” In Proceedings of the 19th international conference on World wide web (ACM), 2010, pp. 891-900.

[23] E. Momeni, C. Cardie and M. Ott, “Properties, Prediction, and Prevalence of Useful User-Generated Comments for Descriptive Annotation of Social Media Objects,” In Proceedings of ICWSM, 2013.

[24] O. Uryupina, B. Plank, A. Severyn, A. Rotondi and A. Moschitti, “SenTube: A Corpus for Sentiment Analysis on YouTube Social Media,” In LREC, 2014, pp. 4244-4249.