Smart Intelligent Computing and Communication Technology
V.D. Ambeth Kumar et al. (Eds.)

doi:10.3233/APC210099

Scraping and Analysing YouTube Trending Videos for BI

Sowmiya K[a,1], Supriya S[b] and R.Subhashini[c]
[a,b]*UG Scholar, Dept of IT, Sathyabama Institute of Science and Technology, Chennai, India*
[c]*Professor, Dept of IT, Sathyabama Institute of Science and Technology, Chennai, India*

Abstract: Analysis of structured data has seen tremendous success in the past. However, large scale of an unstructured data have been analysed in the form of video format remains a challenging area. YouTube, a Google company, has over a billion users and it used to generate billions of views. Since YouTube data is getting created in a very huge amount with a lot of views and with an equally great speed, there is a huge demand to store the data, process the data and carefully study the data in this large amount of it usable. The project utilizes the YouTube Data API (Application Programming Interface) which allows the applications or websites to incorporate functions in which they are used by YouTube application to fetch and view the information. The Google Developers Console which is used to generate an unique access key which is further required to fetch the data from YouTube public channel. Process the data and finally data stored in AWS. This project extracts the meaningful output of which can be used by the management for analysis, these methodologies helpful for business intelligence.

Keywords: Datasets, Data Analysis, Social Media, BigData, Decision Making, Amazon S3, Cloud Computing.

1. Introduction

Now a days social media like Facebook, Twitter, YouTube and Google make the space for millions of users to share their opinion each other. In the rapidly increasing popularity, we have these sites that have become a source of massive amount in the real time data of videos, images etc. Among them, YouTube is one of the world's largest video sharing platforms, where videos are uploading continuously by the millions of users. The YouTube has emerged as a comprehensive and easy to access the compilation of video information source on the web. It is a unique environment with many facets such as multi-modal, multi-lingual, multidomain and multi-cultural. The versatility and attractive shared content draw the widespread attention. Therefore, the importance of YouTube is successively increasing for the industry and research community day by day. Increase interaction of user's , it allows users express their

[1]Sowmiya K,UG Scholar, Dept of IT, Sathyabama Institute of Science andTechnology, Chennai, India
E-mail: sowmiyakumar1998nov@gmail.com

opinion by rating the viewed object's interaction and object with the community Moreover, these data which serves the purpose of helping the community to filter relevant opinions more efficiently. The researchers continuously showing their interest in social media data analysing to exploiting rich content shared on YouTube. In earlier studies the videos not only using meta data, but also using comment. Unstructured formats in YouTube is comments and it's difficult part to analyse. In this paper we used a NLP sentiment analysis approach to find out the YouTube trending video. In this paper, seeking, how metadata useful for analysis. Data analysis is a process of processing, cleaning irrelevant data, transform the data, and modelling goal of discover, informing conclusion and help to promote decision making for business.

2. Relative Work

Several researches have taken from the different regions from the YouTube video methods [5].In that regions from YouTube comments are the most important thing to create or to make a decision (like comment rating, searching the topics related to the categories etc.,) of the particular chosen video [6]. The above searched comments are used to promote the video objects that are taken from the particular chosen video [7, 9].The Comments will also shows us the users behaviour and we could able to find out the most repeated comments and the troll messages about that video or the video maker and we could find the troll makers [1].After analysing the objects of the comments it is easy and it is possible to find out the users view on the v either it is positive or negative, and this is known as the sentimental analysis[10]. And on the basis of the comments we can able to categorize the trending videos in several category to know the users view [11, 12]. On the upcoming process for the improvement we can improve the video in the method based on the basic needs and the feature of the social medias or social networks [3].And have to work on the YouTube trending videos regarding comments for like and dislikes for showing that user's thought that are influenced by the most valuable comments [4]. These two methods is use to find how popular that the video is trending and the categories that are trending , so that it could help us to retrieve the useful video for promoting the advertisements. These two proposed works [3, 4] shows that the interesting work for video and its retrieval process but they used like/dislike and views. But sometimes it may lead to the wrong or inaccurate results.And we analyse the large amount of comments instead of others comment like view and comments to find the related videos which can be useful toyoutube users [11-14].

3. Proposed System

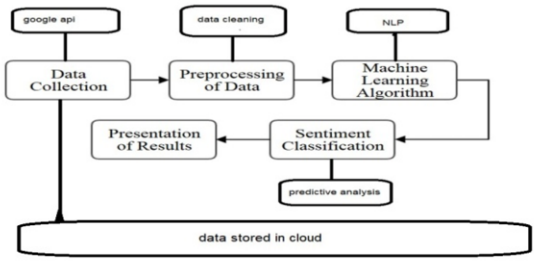

Figure 1. Architectural diagram

3.1 Data Collection

Data collection is the process where we collect the data from the google API for seven different countries and each country has its own country code and the google API is in alpha numerical numbers and it also provides the overview of the functions.

Before startingthe process

- First we should need the Google account to get the Access from the Google API console, then we have to request the key and register for our application.
- Then we have to create or form a project in the Google console and then we have to get the correct permission from that API for our application to submit our request.
- After completing our project, we have to make sure of the data that are collected from the YouTube API is registered for the application to use :
    - a.Firstwe have to go to the Google API console and then we have to select our project that we have registered.
    - b.Then we have to check the form from API page that have been enabled and then we have to check whether the process is ON for the data in the YouTube API.
- The application may use any methods from the API and we have to understand properly that how to implement in that user authorization.
- Then we have to select the client lib to make clear for our progress.
- First we have to make clear about the concepts of the JavaScript Object Notation data format.
- Finally, in Spyder notepad we run the code to get the semi structured data.

3.2 Data Processing

In the Data processing method the types and the methods to convert the normal information or normal messages to the high level knowledgeable and similar to the method in the data analysis process.The data that we are getting first must be processed for that analysis part. These data can involve in substituting or placing those data into rows and into columns in the form of table and it is also known as the structured data and in the future progress the analysis can be used in the form of spreadsheet or in the form of statistical software.

3.3 Data Pre-processing

After the processing method and after we organized the data, the data may be not in the proper form and it also have the chances of containing duplicates of that data and having some error in it. Then we must have to clean the unwanted data like cleaning the unwanted alpha numerical numbers on that data this will form the problems for the data that is stored and entered. And Data cleaning is the process of cleaning the data and removing the errors of that data. And some of the steps to include in the cleaning process is rectifying the inaccurate data and changing the duplicate data and for example, bothtrending date and publish time column include dates, but in different formatting. Publish time includes the time of publication and the date, while trending date only includes the date. Since we don't have the time the video started trending, we cannot compare it to the publish time. In the end, we will clean the data such that both columns include datetime.date objects then We will clean up the tags column by separating tags into a list instead of one long string. NLP process to remove title We developed a function to clean the video titles by removing stop words, non-alphanumeric tokens, and money amounts to better understand the occurrence of certain keywords being used in the titles.

3.4 Exploratory Data Analysis

After the datasets are cleared and cleaned and the next process was analysing part. The analysation part contains many methods and to start with it we can understand the messages which contain within the required data. And the process of exploring the data may give the result as the addition of cleaning the data or the addition of requesting the data. The analysing part will have the graph of showing that the frequency of the trending videos. And the next process will be Data visualisation process.

3.5 Data Visualization

Successfully cleaned our data, we will start exploring our data. We will start by looking at what words are commonly used in video titles, as we predicted that word choice will be an important factor for Trending videos. Figure 2, help to understand why top 10 channels with trending videos. may not coincide with the exploration of categories we did earlier as the distribution is extremely rightly skewed. Around 1450 YouTube channels have under 10 YouTube trending videos, while we have few trending channels over with 100 YouTube trending videos. Theses Visualization tells ,look too closely the top 10 YouTube channels won't help us as the majority we have below, that number of YouTube videos. Calculate the percentages of the number of YouTube trending videos theses channels have to focus more closely.

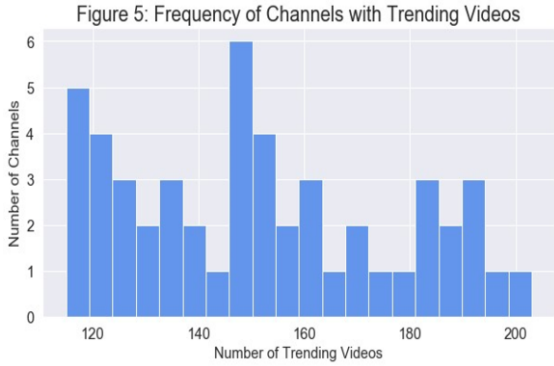

Figure 2. Number of trending videos by channe1 Percentage.

Percentage = (part / total) × 100

above_20 =len([i**for**i**in**channel_freq**if**i> 20]) /len(channel_freq) * 100 percent_16_20 =len([i**for**i**in**channel_fre q **if** (i<= 20) **&** (i> 15)]) /len(channel_fre q) * 100

percent_11_15 =len([i**for**i**in**channel_fre q **if** (i<= 15) **&** (i> 10)]) /len(channel_fre q) * 100

percent_6_10 =len([i**for**i**in**channel_freq**if** (i<= 10) **&** (i> 5)]) /len(channel_freq) *

100 percent_1_5 =len([i**for**i**in**channel_freq**i f** (i<= 5)]) /len(channel_freq) * 100

Number of Trending Videos by Channel

- 1 to 5->Videos -34.5%
- 5 to 11->Videos - 24.8%
- 11 to 5->Videos -10.8%
- 15 to 20->Videos -7.1%
- 21+ -> Videos - 22.9%

This demo, that more than 3/4 of channels have created under 20 YouTubetrending videos, with 34.5 percentage of channels having between 1 - 5YouTubetrending videos. This shows the top channels with well over 100 trending videos.

3.6 Stored In Amazon S3

Amazon Simple Storage Serviceespecially formulated to stored Semi-structured data and Unstructured data.

1. Get starting with AWS management console and move to S3.
2. Creating bucket name is socialmd**in 'US East (Ohio) us-east-2'** region and stored the objects. (every object will have unique URL link)

Once objects have been stored in an 'Amazon S3 bucket'; they are given an **object key**. Use this, along with the bucket name, to access the object. An **object key** is the

unique identifier for which an object in a bucket. Because the combination of a "bucket", "key", and "version ID" identifies uniquely each object, you can think of 'Amazon S3' as the basic data map between bucket + key + version andobject. In each object in'Amazon S3' can be uniquely called (or) addressed through the combination of the web service –'endpoint', 'bucket name','key', andoptionally (version).
'Amazon S3' is designed 11 9'sdurability, and its encrypted.

4. Conclusion

This paper illustrates to get the right decision for digital marketing for company growth, especially marketer can market they product using smart analytical processes. YouTube trending video helps businesses understand how customers feel about their brand, giving them first-hand information to improve their products, make data-driven decisions, and deliver better customer experiences.

Reference

[1] A.Severyn,A.A.Moschitt.Uryupina, B.Plank and K. Filippo, Multi-lingual opinion mining on YouTube,"Information Processing & Management, 52(1), 2016, pp. 46-60.
[2] YouTube data API documentation https://developers.google.com/youtube/2.0/reference
[3] Cholera.S, 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.
[4] Schulte's.P, V. Dorner and F. Lehner, Leave a Comment! An In-depth Analysis of User Comments on YouTube, Wirtschaftsinformatik, 2013, pp. 659-673.
[5] Siersdorfer.S, S. Cholera, J. S. Pedro, I. S. Altingovde and W. Nejdl, Analysing and mining comments and comment ratings on the social web, ACM Transactions on the Web (TWEB), 8(3), 2014, pp. 1-39.
[6] Siersdorfer.S, S. Cholera, W. Nejdl and J. San Pedro, How useful are your comments? analysing and predicting YouTube comments and comment ratings, In Proceedings of the 19th international conference on World wide web (ACM), 2010, pp. 891-900.
[7] Mokena.E, 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.
[8] https://www.kaggle.com/ammar111/yo utube-trending-videos-analysis
[9] H.Lee,Yehan, Y . Kim and K. Kim, Sentiment analysis on online social network using probability Model, In Proceedings of the Sixth International Conference on Advances in Future Internet, 2014, pp. 14-19.
[10] Filippova.K and K. B. Hall. Improved video categorization from text metadata and user comments. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, 2011, pp. 835-842.
[11] Ambeth Kumar.V.D , Ramya.S , Divakar.H , KumuthaRajeswari.G .A Survey on Face Recognition in Video Surveillance. Lecturer Notes on Computational and Mechanism, Vol. 30, pp: 699-708, 2019
[12] Ambeth Kumar.V.D .Precautionary measures for accidents due to mobile phone using IOT.Clinical eHealth, Volume 1, Issue 1, March 2018, Pages 30-35.
[13] Nanagasabapathy.K et.al . Validation system using smartphone luminescence. IEEE International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Pages: 235 – 239, 6-7 July 2017, Kannur, India
[14] Ambeth Kumar.V.D and Dr.M.Ramakrishan, "Enhancement in Footprint Image using Diverse Filtering Technique", Procedia Engineering journal, Volume 8, No.12, 1072-1080, March 2012 . [doi:10.1016/j.proeng.2012.01.965]