ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 5, Issue 6, June 2018

Classification of YouTube Data based on Sentiment Analysis

[1] Shaila S.G, [2] Prasanna MSM, [3] Kishore Mohit

[1][2][3] Dayananda Sagar University, Bangalore, India.

Abstract- **Nowadays, Big data and Data mining have attracted a great deal of attention in the information industry, due to the wide availability of huge amounts of data and the urgent need for turning such data into useful knowledge through predictive models. Corporate companies are using social media for improving their businesses, the data mining and analysis are very important in these days. Thus, Interaction and review are established with the customers and the concept, characteristics & need for Big Data & different offerings available in the market to explore unstructured large data. The paper deals with analysis of YouTube Data. The analysis is done using users Sentiments features such as Views, Comments, Likes, and Dislikes. We used the Linear Regression classification approach to classify the YouTube Data. The experimental results are given accurate results which illustrated that it is influential practice and a key enabler for the social business. The insights gained from the user generated online contents and collaboration with customers is critical for success in the age of social media.**

I. INTRODUCTION

With rapid innovations and surge of internet companies like Google, Yahoo, Amazon, eBay and a rapidly growing internet savvy population, today's advanced systems and enterprises are generating data in a very huge volume with great velocity and in a multi-structured format including videos, images, sensor data, weblogs etc. from different sources. This has given birth to a new type of data called Big Data which is unstructured sometime semi structured and also unpredictable in nature. This data is mostly generated in real time from social media websites which is increasing exponentially on a daily basis. According to Wikipedia, "Big Data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Analysis of data sets can find new correlations to spot business trends, prevent diseases, combat crime and so on [1]". With millions of people using Twitter to tweet about their most recent brand experience or hundreds of thousands of check-ins on Yelp, thousands of people talking about a recently released movie on Facebook and millions of views on YouTube for a recently released movie trailer, we are at a stage wherein we are heading into a social media data explosion. Companies are already facing challenges getting useful information from the transactional data from their customers for instance, data captured by the companies when a customer opens a new account or sign up for a credit card or a service. This type of data is structural in nature and still manageable. However, social media data is primarily unstructured in nature. The very unstructured nature of the data makes it very hard to analyze and very interesting at the same time. Whereas RDBMS is designed to handle structured data and that to only certain limit, RDBMS fails to handle this kind of unstructured and huge amount of data called Big Data. This inability of RDBMS has given birth to new database management system called NoSQL management system. YouTube is a open source platform from google and have billion users that are present and everyday people watch hundreds of millions of hours on YouTube and that generate billions of views. Every day, people across the world are uploading 1.2 million videos to YouTube, or over 100 hours per minute and this number is ever increasing [3]. To analyze and understand the activity occurring on such a massive scale, a relational SQL database is not enough. Such kind of data is well suited to a massively parallel and distributed system like Hadoop..

II. LITERATURE REVIEW

Social media has generated a wealth of data. Billions of people tweet, sharing, post, and discuss every day. Due to this increased activity, social media platforms provide new opportunities for research about human behavior, information diffusion, and influence propagation at a scale that is otherwise impossible. Social media data is a new treasure trove for data mining and predictive analytics. Since social media data differs from conventional data, it is imperative to study its unique characteristics. Big Data in technology has contributed much toward marketing practitioners to search for new patterns when faced with assessing brand performance during brand equity appraisal. Some challenges of current practices are that these methods rely heavily on data collection and analysis methods such as question, which have a significant time lag. In the paper, authors (Jai Prakash Verma and Giaglis, 2017) has defined a computational model that add up the topic and sentiment classification subjects from consumer point of view in social media network. Their model devises a novel algorithm to improve clustering of post in semantically related groups, which is an essential prerequisite when searching for related topics and sentiment in pools of big data. To define the validity of their algorithm, they apply it to the other transportation network, from data collected through Twitter or any other social media. Results obtained present consumer sentiment towards data and produce insights for fundamental brand equity dimensions. The authors in the work (Morstatter and Liu, 2017) propose computational methods to assess if there is bias due to the way a social media site makes its data available, to detect bias from data samples without access to the full data, and to mitigate bias by designing data collection strategies that maximize coverage to minimize bias. They also present a new kind of data bias stemming from API attacks with both algorithms, data, and validation results. This work demonstrates how some characteristics of social media data can be extensively studied and verified and how corresponding intervention mechanisms can be designed to overcome negative effects.

III. PROPOSED APPROACH

The approach focuses on how data generated from YouTube can be mined and utilized by different companies to make targeted, real time and informed decisions about their product that can increase their market share. The approach uses Hadoop concepts. The approach has multiple applications. Industries utilize the approach to understand the effectiveness of marketing. In addition to the view counts, subscribers and shares, audience retention count, companies can also evaluate views according to date range. This can tell the industries when the slow period or spike in viewership is and attribute the same to certain marketing campaign. Applications for YouTube data can be endless. For example, industries can analyze how much a product is liked by people. The application helps in analyzing new emerging trends and knowing about people's changing behavior with time. Also, people in different countries have different preferences. By analyzing the comments/feedbacks/likes/view counts etc.

The approach uses several steps and techniques to identify the optimized algorithm for classification of YouTube Data. The following are listed below:

- Fetching YouTube Data Using Google API
- Storing Data to Hadoop Distributed File System (HDFS)
- Indexing Data
- Cleaning Data
- Creating Model for Data
- Analysis of Data
- Training Data
- Classification of Data by selecting best Fitting Algorithm for Social media Analysis: YouTube

3.1 Fetching YouTube Data Using Google API

The Fig 1 below presents YouTube Extracted Data Using Google API.

--728h8mnDY,2017-12-07,Rösti - YouCook,YouCook Cuisine,26, Howto & Style,2017-12-05T16:30:00.000Z,39535,2573,51,269,False,False,False,FR,0.980564024390243 9,
-1446KeS7aA,2017-12-23,Servietten falten Weihnachten: Schneeflocken basteln mit Papier-Servietten - DIY Origami Deko,Basteln mit Papier,26, Howto & Style,2017-12-22T10:
-2lXJX54cPU,2017-12-19,Schönen 3. Advent,eKartenwelt.de,26, Howto & Style,2017-12-17T06:20:04.000Z,27361,0,0,0,True,True,False,DE,,,
-77ZOJ7sjpY,2017-12-14,PARTY LOOKBOOK : TENUES DE SOIREE (SPECIAL FÊTES),Mimi Ar,26, Howto & Style,2017-12-13T17:34:46.000Z,12665,909,34,95,False,Fa
-7uBsC4maVA,2017-12-17,DIY NOEL JE TEST 3 DIY PINTEREST - DECO FACILES ou PAS français,By Isnata,26, Howto & Style,2017-12-16T15:29:46.000Z,15587,1
-A_XQW360wc,2017-12-11,Masha Birthday Cake / Geburtstagstorte / Mascha und der Bär/ Маша и медведь,Sallys Welt,26, Howto & Style,2017-12-10T09:41:03.000Z,83709,
-CqiTME15Ms,2017-11-26,3 DIY À OFFRIR SUPER FACILES ! avec The Doll Beauty et Perfect Honesty,Rose Carpet,26, Howto & Style,2017-11-25T10:00:09.000Z,173355,1
-JzB-IP_hok,2017-12-03,Coupez un oignon en 4 tranches et répartissez-les dans votre maison. La raison ? - France 365,FRANCE 365,26, Howto & Style,2017-12-02T15:12:53.0
-OFyzARkKbw,2017-12-04,شوفو تحضيرات عيد ميلادي في شحال فعمري ؟؟ فتح الهدايا أفكار بسيطة لعيد الميلاد لا تفوتكم!!,Malak World - 26,عالم ملاك, Howto & Style,2017-12-03T09:00:01.000Z
-PugqHo4n3w,2017-12-09,Vergrabe ein Ei im Boden deines Gartens - Was ein paar Tage später passiert wird dich überraschen!,Gesundheitsblatt,26, Howto & Style,2017-12-08T1
-Wwv_5QBwfs,2017-11-20,Tasty 101: How To Cook Perfect Chicken Breast Every Time,Tasty,26, Howto & Style,2017-11-19T16:00:18.000Z,144002,6047,77,446,False,False,
-b2C0M8afgg,2017-11-21,à tous les gens meurtris,ELO,26, Howto & Style,2017-11-20T16:56:48.000Z,13486,2839,13,207,False,False,False,FR,0.995441795231416 5,0.0045
-cukFZbuj8g,2017-12-16,DIY: beleuchtetes Silvester-Dekoschild für 2018 [How to] Deko Kitchen,Deko Kitchen – Schöne Deko selber machen | Esther Straub,26, Howto & Style,
-ew7iObcQiU,2017-12-04,20 IDÉES BRILLANTES QUI TE RENDRONT LA VIE PLUS FACILE,BRICO SYMPA,26, Howto & Style,2017-12-03T15:00:01.000Z,4230,129,2,
-guI-Q8jYH0,2018-01-03,4 Astuces avec Préservatif VOUS DEVRIEZ SAVOIR,Mr. Henry,26, Howto & Style,2018-01-02T16:00:00.000Z,2663,21,6,0,False,False,False,FR,0.7
-m0lHGq1OKs,2017-12-20,La recette facile de Tarte Tatin aux pommes caramélisées,Hervé Cuisine,26, Howto & Style,2017-12-19T07:30:02.000Z,12414,1104,19,89,False,Fals
-pyLfpYxmAk,2017-12-30,Guten Rutsch / Silvestergrüße,eKartenwelt.de,26, Howto & Style,2017-12-29T05:30:55.000Z,10262,0,0,0,True,True,False,DE,,,
-sWdRXuRh0E,2017-12-17,FUITE DE DOCUMENT! Comment la POLICE NATIONALE va-t-elle CONTROLER LES DRONES en 2018?,Alex Cipriani Travel,26, Howto & S
-tjucPZlcgc,2018-01-05,F&F Active | #whatsstoppingyou,F&F Clothing,26, Howto & Style,2018-01-01T11:34:32.000Z,2414,5,43,9,False,False,False,GB,0.10416666666666
-tqKJE4v-nE,2017-12-17,Chocolate Bark - Holiday Gift Idea - Food Wishes,Food Wishes,26, Howto & Style,2017-12-16T01:50:41.000Z,98126,4415,45,354,False,False,False,
03RQLmDNIEw,2017-12-05,Flying Home For Christmas | Vlogmas Days 1 & 2,Tanya Burr,26, Howto & Style,2017-12-03T15:00:01.000Z,243420,10374,292,387,False,False,
03RQLmDNIEw,2017-12-11,Flying Home For Christmas | Vlogmas Days 1 & 2,Tanya Burr,26, Howto & Style,2017-12-03T15:00:01.000Z,330158,12536,388,292,False,False,
03gu7d9DDOk,2018-01-03,DIY Barbie Doll Hairstyles Barbie Hair Transformations,Hairstyle Tutorials,26, Howto & Style,2018-01-02T16:41:52.000Z,54940,760,45,66,Fa
0ANDiPOPS_O,2017-11-27,♡ SILVER PLAYBUTTON UNBOXING (and making it cute) ♡,Pixielocks,26, Howto & Style,2017-11-15T15:17:28.000Z,28869,3110,90,454,False

Fig 1. YouTube Extracted Data Using Google API

3.2 Storing Data to Hadoop Distributed File System (HDFS)

The following steps are followed to store data to Hadoop Distributed File System (HDFS). First, we need to login to Cloudera cluster with the login credentials provided by Cloudera.com. Next FTS server is used to upload dataset to Local drive so it will be available all the time if anything goes wrong in future, that requires only upload features to upload the dataset to local file system of Cloudera. Later, once file has been uploaded successfully, need to store data to HDFS so that MapReduce job to Hadoop cluster can be done. Hadoop commands are used to copy the dataset stored at local server to Hadoop cluster as shown below.

$hadoop fs -put youtube.txt youtube.txt

Or, can be copied data directly to HDFS using Hue Framework provided by Cloudera just like Upload and download the data. This completes uploading dataset to HDFS for future analysis.

3.3 Indexing Data

Hive Query is used for Indexing as follows and represented in Fig 2.

create table YouTube (VideoId string, Trending_Date Date, title string, channel_title string,category_id int,Category_name string, publish_time int, views bigint, likes bigint, dislikes bigint, comment bigint,comment_disable boolean, ratings_disable boolean,video_error_or_removed boolean, country string,likes_prop float, dislikes_prop float,like_ratio float row format delimited fields terminated by '\t' location 'hdfs://nameservice1 /user /edureka_ 359332 /Youtubbe';

	A	B	C	D	E	F	G	H	I	J
1	video_id	trending_date	title	channel_title	category_id	category_name	publish_time	views	likes	dislike
2	--728h8mı	12/7/2017	RÃ¶sti - Yı	YouCook Cuisine	26	Howto & Style	2017-12-05T16	39535	2573	
3	-1446KeS7	12/23/2017	Serviettei	Basteln mit Papi	26	Howto & Style	2017-12-22T10	12603	304	
4	-2lXJX54cl	12/19/2017	SchÃ¶nen	eKartenwelt.de	26	Howto & Style	2017-12-17T06	27361	0	
5	-77ZOJ7sjı	12/14/2017	PARTY LO(	Mimi Ar	26	Howto & Style	2017-12-13T17	12665	909	
6	-7uBsC4m	12/17/2017	DIY NOEL (	By Isnata	26	Howto & Style	2017-12-16T15	15587	1131	
7	#NAME?	12/11/2017	Masha Bir	Sallys Welt	26	Howto & Style	2017-12-10T09	83709	3287	
8	#NAME?	11/26/2017	3 DIY Ã€ O	Rose Carpet	26	Howto & Style	2017-11-25T10	173355	11518	
9	#NAME?	12/3/2017	Coupez ur	FRANCE 365	26	Howto & Style	2017-12-02T15	6430	88	
10	#NAME?	12/4/2017	Ø´Ù ÙÙ ç	Malak World - Ø	26	Howto & Style	2017-12-03T09	86184	6984	
11	#NAME?	12/9/2017	Vergrabe (	Gesundheitsblat	26	Howto & Style	2017-12-08T19	10789	768	
12	#NAME?	11/20/2017	Tasty 101:	Tasty	26	Howto & Style	2017-11-19T16	144002	6047	
13	#NAME?	11/21/2017	aì€ tous le	ELO	26	Howto & Style	2017-11-20T16	13486	2839	
14	#NAME?	12/16/2017	DIY: beleu	Deko Kitchen â€	26	Howto & Style	2017-12-15T14	5500	212	
15	#NAME?	12/4/2017	20 IDÃ‰E	BRICO SYMPA	26	Howto & Style	2017-12-03T15	4230	129	
16	#NAME?	1/3/2018	4 Astuces	Mr. Henry	26	Howto & Style	2018-01-02T16	2663	21	
17	#NAME?	12/20/2017	La recette	HervÃ© Cuisine	26	Howto & Style	2017-12-19T07	12414	1104	
18	#NAME?	12/30/2017	Guten Rut	eKartenwelt.de	26	Howto & Style	2017-12-29T05	10262	0	
19	#NAME?	12/17/2017	FUITE DE [	Alex Cipriani Tra	26	Howto & Style	2017-12-16T09	5415	324	
20	#NAME?	1/5/2018	F&F Active	F&F Clothing	26	Howto & Style	2018-01-01T11	2414	5	
21	#NAME?	12/17/2017	Chocolate	Food Wishes	26	Howto & Style	2017-12-16T01	98126	4415	
22	03RQLmDI	12/5/2017	Flying Hor	Tanya Burr	26	Howto & Style	2017-12-03T15	243420	10374	
23	03RQLmDI	12/11/2017	Flying Hor	Tanya Burr	26	Howto & Style	2017-12-03T15	330158	12536	
24	03gu7d9D	1/3/2018	DIY Barbie	Hairstyle Tutoria	26	Howto & Style	2018-01-02T16	54940	760	
25	0ANDjPQI	11/27/2017	â™¡ SILVEI	Pixielocks	26	Howto & Style	2017-11-15T15	28869	3110	
26	0ANDjPQI	11/20/2017	â™¡ SILVEI	Pixielocks	26	Howto & Style	2017-11-15T15	27408	2993	
27	0EGndnte	12/3/2017	Zentralsta	Sallys Welt	26	Howto & Style	2017-12-02T08	131939	4314	
28	0Fo5rISAJ	11/22/2017	25 INGENI	5-Minute Crafts	26	Howto & Style	2017-11-21T08	897452	7957	
29	0GooTyuP	12/1/2017	Ø¬Ù†Ø³Ù	Malak World - Ø	26	Howto & Style	2017-11-30T08	86818	6164	
30	0MUCYFLI	11/26/2017	â™¡â€¢ D(	L'atelier de Roxa	26	Howto & Style	2017-11-24T16	233420	31847	
31	0N0gjw4C	12/24/2017	NEW FENT	LustreLux	26	Howto & Style	2017-12-23T21	72125	6296	
32	0Rt329xQI	11/24/2017	Brad and S	Bon AppÃ©tit	26	Howto & Style	2017-11-21T16	289999	10239	
33	0Rt329xQI	11/29/2017	Brad and S	Bon AppÃ©tit	26	Howto & Style	2017-11-21T16	352742	11503	
34	0TKzJZQc6	12/10/2017	Tourtiere	Food Wishes	26	Howto & Style	2017-12-09T00	140696	7216	
35	0Uq8EZzf1	12/19/2017	KRANSEK/	Carl is cooking	26	Howto & Style	2017-12-18T15	9318	954	
36	0W16_CO	12/28/2017	FAVORITE	thataylaa	26	Howto & Style	2017-12-28T02	40022	4412	
37	0_j08BUAI	12/2/2017	DIY: Weih	Deko Kitchen â€	26	Howto & Style	2017-12-01T14	4884	438	
38	0asPIGtVi	11/27/2017	Tasty 101:	Tasty	26	Howto & Style	2017-11-26T16	169794	5780	
39	0bjPjYYT1I	12/21/2017	Die bestei	PULS Reportage	26	Howto & Style	2017-12-20T14	83783	3678	
40	0dAwMQ3	12/3/2017	Weirdest	Clevver Style	26	Howto & Style	2017-12-02T16	108648	5592	

Fig 2. Indexed Data created from YouTube Dataset

Finding other Insights in data Using MapReduce Java Program and Hive Query for better understanding of data for Machine Learning section is used and is represented in Table 1 and 2 and Fig 3 and 4.

Hive Query1
select videoid, uploader, category, rating from youtube ORDER BY rating desc limit 10;

Query History | Saved Queries | Results (10)

	videoid	uploader	category	rating
1	geUY_esOrt0	samsarajade	Comedy	5
2	osqnkdc349s	giantPANDER	News & Politics	5
3	wzUyV42tzz4	ELmaDDDin	Entertainment	5
4	dh6dF1XY3ul	Ffllaanndd	Entertainment	5
5	L3mR8syHNlg	DarkArchon102	Entertainment	5
6	YZev1imoxX8	threefirstnames	Entertainment	5
7	gP0jnBrVEpl	NoodleStroodel	Film & Animation	5
8	Y_hAXbgxXp8	andiKentxmai	News & Politics	5
9	lD3GuCoJxT4	leetvideos	Education	5
10	3TYqkBJ9YRk	hggh22	Comedy	5

Table 1: You tube data order by Rating

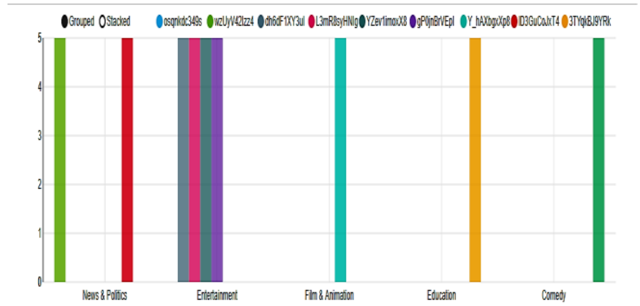

Fig 3. Hive Query Output for Top 10 Rated videos on Youtube

Hive Query2
select videoid, uploader, category, views from youtube order by views desc limit 10;

	videoid	uploader	category	views
1	12Z3J1uzd0Q	kaejane	Film & Animation	65341925
2	4DC4Rb9quKk	ChrisBrownTV	Music	33754615
3	LU8DDYz68kM	Jason275	Pets & Animals	27721690
4	kHmvkRoEowc	itschriscrocker	Entertainment	18235463
5	Md6rURKhZmA	TPainVideos	Music	18141492
6	EwTZ2xpQwpA	TayZonday	Music	16841569
7	A2f3cuUXXRs	RKellyTV	Music	13038204
8	rZBA0SKmQy8	TownIdiot25	Comedy	11007201
9	irp8CNj9qBI	Frozentoast	Music	10172172
10	ZCYaw5tGYAs	nozzle49	Comedy	8944331

Table 2: You tube data order by Views

➢ format = "%Y/%m/%d")

• Now check the missing data in dataset using Model if there is any missing value in the data it will represent using yellow color in Heatmap of data.

➢ sns.heatmap(youtube.isnull(),yticklabels=False,cbar=False,cmap='viridis')

The missing data from data set is represented in Fig 5 which is processed for cleaning as shown in Fig 5.

• Next, cleaning is performed on missing data, some data are neglected as it consist of redundant and has noisy values. These data cause much impact.

➢ Youtube = youtube.dropna()

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 6, June 2018

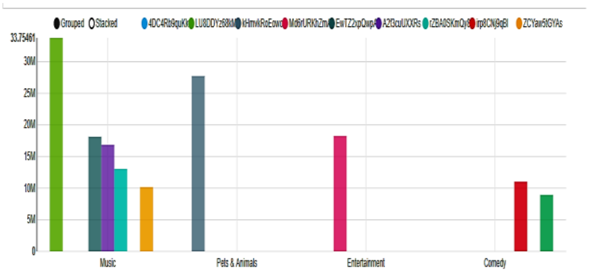

Fig 4. Hive Query output for top 10 categorical views on video.

3.4 Cleaning Data

After indexing of dataset, we need to process the data for data cleaning to remove redundancies and to fill up missing data. Following are the steps:

- To load dataset for cleaning the data using libraries
  - %matplotlib inline
  - import pandas as pd
  - import matplotlib.pyplot as plt
  - import numpy as np
  - import seaborn as sns
  - from datetime import datetime

- Then read data as youtube data
  - youtube = pd.read_csv("youtubeVid_main.csv",sep = ",")
  - youtube["trending_date"] = pd.to_datetime(youtube["trending_date"] \,

Out[46]: <matplotlib.axes._subplots.AxesSubplot at 0x252bbb93748>

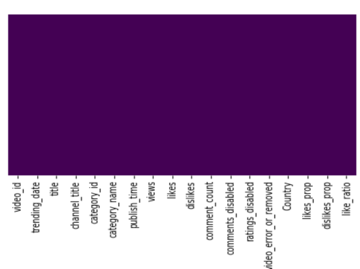

Fig 5. YouTube Data after performing cleaning

3.5 Classification of YouTube Data

Now, data is cleaned and ready for machine learning. We used three classification algorithms for classifying YouTube data based on users sentiment analysis.

- Logistic regression
- Decision Tree
- Random Forest

We need to identify how much data is linearly correlated to each other hence some analysis is needed to find correlation.

IV. EXPERIMENTAL RESULTS

We correlate the YouTube data depending on Views, Likes and Dislikes. To do this we need to know weather data is linear. The following code is used to know the correlation.

```
corr = Youtube.corr()
plt. Figure (figsize= (12,8))
sns. Heatmap (corr, annot=True, cmap='rainbow')
```

Fig 6 represents Correlation generated between data attributes. We can visualize that every column is correlated to each other. We can also visualize that Views, Likes, Dislikes, Comments are correlated with 85% approximately.

<matplotlib.axes._subplots.AxesSubplot at 0x252bc709e10>

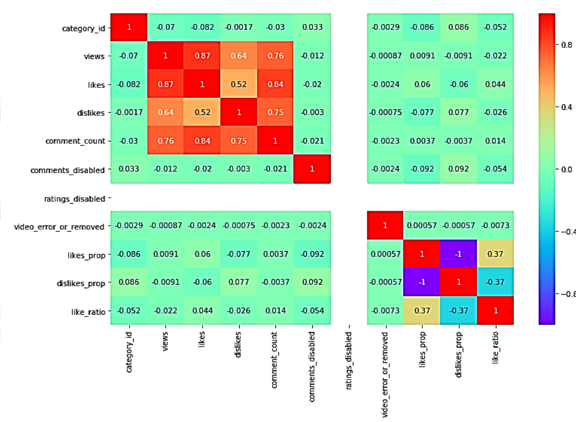

Fig 6. Correlation with respect to Data attributes Views, Likes, Dislikes, Comments

Top 10 YouTube channels can be extracted using the following code and is represented in Fig 7.

```
topUsers_df=pivotTab_toDF(dataset=youtube,value="views", indices = "channel_title", aggregFunct = np.sum)[:10]
print ("Top 10 Youtube Channels:\n")
topUsers_df
```

Out[28]:

	channel_title	total_views
0	YouTube Spotlight	674196442
1	Ed Sheeran	558250009
2	LuisFonsiVEVO	387989509
3	Marvel Entertainment	380297875
4	EminemVEVO	263301085
5	GEazyMusicVEVO	224006918
6	5-Minute Crafts	200769312
7	jypentertainment	193145941
8	ibighit	190216895
9	Universal Pictures	186855788

Fig 7 Top 10 YouTube Channels

Top 10 YouTube categories can be extracted using the following code and is represented in Fig 8.

```
topCat_df=pivotTab_toDF     (dataset=youtube,
value="views",       \indices="category_name",
aggregFunct= np.sum) [:5]
print ("Top 5 Youtube Categories:\n")
topCat_df
```

Out[30]:

	category_name	total_views
0	Music	4740513137
1	Entertainment	4735478990
2	Comedy	1544057914
3	People & Blogs	1031071829
4	Sports	921962563

Figure 8. Top 5 Youtube Category

The classified data is shown in the form of graph which exhibits the trend videos based on Likes and Dislike relation to their Views & Comments.

```
sns.pairplot(df_videos,x_vars=['comment_count
','views'],y_vars=['likes','dislikes'],size=5
)
```

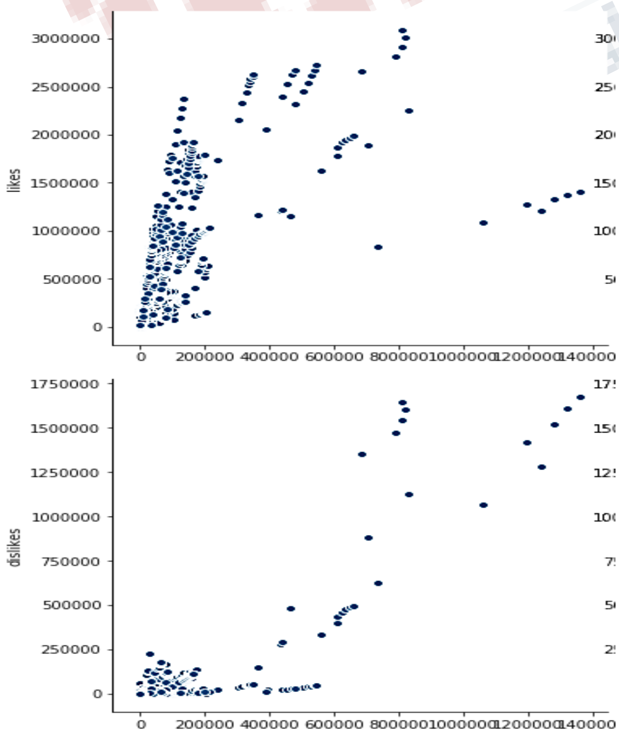

Fig 9. Likes and Dislikes based on Comments and Views

Now, we estimated the user trends based on YouTube data output which is correlated with Views with their Likes ratio and is represented in Fig 10. Below is the following code snippet.

```
sns.jointplot(x='views',y='likes',data=df_vid
eos,kind='reg')
Out[155]: <seaborn.axisgrid.JointGrid at 0x25280c0fc18>
```

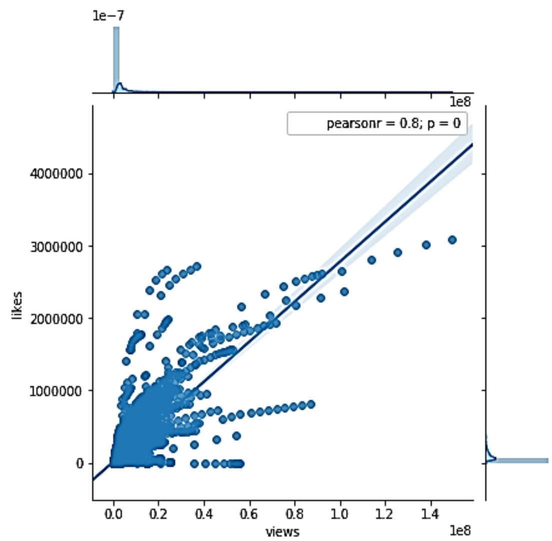

Fig 10. Relation between Likes and Views

The data in Fig 10 is linearly related we can further proceed with linear regression algorithm for classifying data.

3.5.1 Linear Regression for classifying YouTube data

After analyzing, it is known that data is linear correlated in a greater ratio. This can be represented using pair plot of matplotlib.

```
youtube.columnsIndex(['Avg. Likes', 'Comment
Ratio', 'Subscrition', 'Avg. Dislike Ratio',
'Channel          Subscribers',          'Likes',
'Channel'],dtype='object')
sns.pairplot(youtube)
```

After the implementation of the above code we can analyze the relation between attributes of the data. Here, analysis of average Likes, Comment Ratio, Subscrition Avg, Dislike Ratio, Channel Subscribers, Likes Channel, for better understanding. This is shown in Fig 11.

IFERP
connecting engineers... developing research

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 6, June 2018

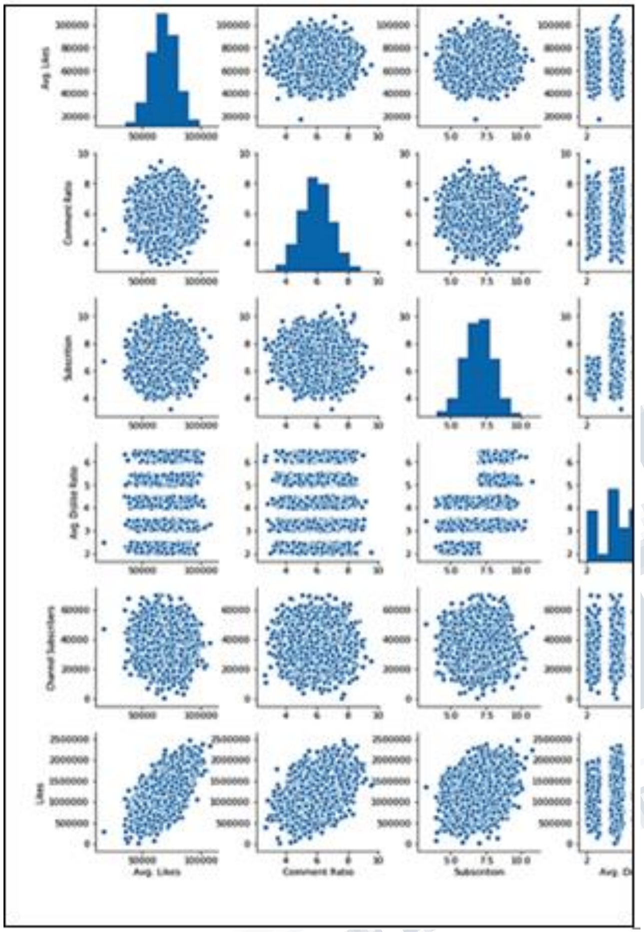

Figure 11 Classification result of Linear Regression

V. CONCLUSION

With the advent of opinion-rich resources such as Social Networks via hand-held devices, people across globe speak out their opinions more open and often. This development has posed us greater opportunities combined with new challenges as people now actively use information technologies to seek out and understand the opinions of others. The insights gained from the user generated online contents and collaboration with customers is critical for success in the age of social media. In this paper, we have demonstrated how Big Data techniques could be used to analyze people's sentiments based on their feedback on YouTube videos. The sample data was extracted using Google's APIs and then was stored into HDFS. Hive was then used to store the datasets in columnar format and query the same to seek the information that is of interest to us.

REFERENCES

[1] CHENG Xu et.al, “Understanding the YouTube and their data”, [IEEE,2014]

[2] Hanif Bhuiyan et.al, “YouTube Video by Sentiment Analysis on User Comment.”, [IEEE, 2017]

[3] Yiheng Liang et.al, “Data Analysis using Hadoop Environment”, [IEEE, 2017]

[4] Bogdan Batrinca et.al, “Social media analytics: a survey of techniques, tools and platforms”, [Springer,2015]

[5] Erik Cambria, “Big Social Data Analysis”, [IEEE,2014]

[6] Jai Prakash Verma et.al, “BIG DATA ANALYTICS: Application for Social Media Data”, [IJSCAI,2016]

[7] Morstatter, F. and Liu H., “Mitigating Data bias in Social Media. Online Social Networks and Media.”.

[8] Mostafa, M.M., “More Than Words: Social Networks' Text Mining for Consumer Brand Sentiments. Expert Systems with Applications”, 40, Issue 10, 4241–4251. [2013]

[9] Naylor, R.W., Lamberton, C.P. and P.M. West, “Beyond the “Like” Button: The Impact of Mere Virtual Presence on Brand Evaluations. Journal of Marketing, 76, No. 6, 105-120.”[2012]

[10] Noone B.M., McGuire, K.A. and K.V. Rohlfs, 2011, Social Media Meets Hotel Revenue Management: Opportunities, Issues and Unanswered Questions. Journal of Revenue and Pricing Management, 10, Issue 4, 293–305.