YOUTUBE TRENDING PREDICTION

Professor: Guiping Hu

IE 587 Big Data Analytics

Submitted by: Smrithi Ajit & Mriga Kher

INTRODUCTION

- YouTube has over 400 hours of content uploaded every minute.
- Content creators and “YouTube” itself gain popularity and monetary benefits when their videos go “viral” or in other words “they trend”.
- What type of videos gain popularity and which ones don’t? The popularity of online videos is typically measured by the number of views they attract from viewers over a short period of time. In this project, we aim to analyze past YouTube dataset and utilize a combination of user attributes to understand their influence on “the popularity”.
- Our target variable that determines the popularity is the Likes Ratio

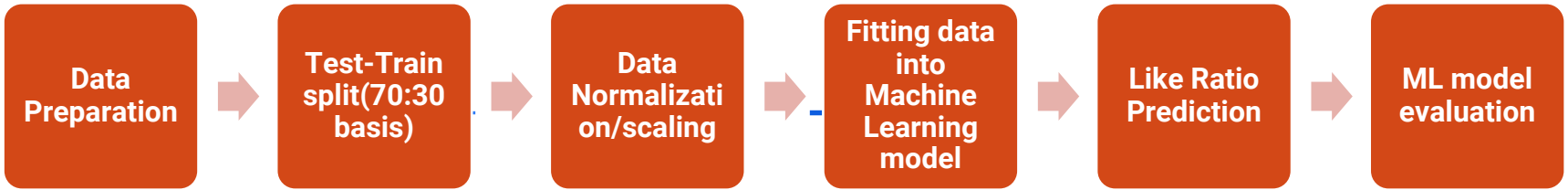
DATA DESCRIPTION

- This dataset includes several months of data on daily trending YouTube videos. Data is included for the USA, Great Britain, Germany, Canada, and France, respectively, with up to 200 listed trending videos per day.
- Only using US dataset (Observations: 45,113, Variables: 16)
- Each region's data is in a separate file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.

SNAPSHOT OF THE DATASET

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	video_id	trending_date	title	channel_title	category_	publish_t	tags	views	likes	dislikes	comment	thumbna	comment	ratings_d	video_er	description		
2	n1WpP7iowLc	17.14.11	Eminem - Walk On Water (Audi	EminemVEVO	10	2017-11-1	Eminem|"'	17158579	787425	43420	125882	https://i.yt	FALSE	FALSE	FALSE	Eminem's new track Walk on Water		
3	0dBIkQ4Mz1M	17.14.11	PLUSH - Bad Unboxing Fan Mail	iDubbbzTV	23	2017-11-1	plush|"bac	1014651	127794	1688	13030	https://i.yt	FALSE	FALSE	FALSE	STill got a lot of packages. Probably		
4	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancu	Rudy Mancuso	23	2017-11-1	racist supe	3191434	146035	5339	8181	https://i.yt	FALSE	FALSE	FALSE	WATCH MY PREVIOUS VIDEO â–¶ \n		
5	d380meD0W0M	17.14.11	I Dare You: GOING BALD!?	nigahiga	24	2017-11-1	ryan|"higa	2095828	132239	1989	17518	https://i.yt	FALSE	FALSE	FALSE	I know it's been a while since we did		
6	2Vv-BfVoq4g	17.14.11	Ed Sheeran - Perfect (Official M	Ed Sheeran	10	2017-11-0	edsheeran	33523622	1634130	21082	85067	https://i.yt	FALSE	FALSE	FALSE	ðŸŽ§: https://ad.gt/yt-perfect\nðŸ":		
7	0yIWz1XEeyc	17.14.11	Jake Paul Says Alissa Violet CHE	DramaAlert	25	2017-11-1	#DramaAle	1309699	103755	4613	12143	https://i.yt	FALSE	FALSE	FALSE	â–¼ Follow for News! - https://twitte		
8	_uM5kFfkhB8	17.14.11	Vanoss Superhero School - New	VanossGaming	23	2017-11-1	Funny Mor	2987945	187464	9850	26629	https://i.yt	FALSE	FALSE	FALSE	Vanoss Merch Shop: https://vanoss.:		
9	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OU	CaseyNeistat	22	2017-11-1	SHANtell n	748374	57534	2967	15959	https://i.yt	FALSE	FALSE	FALSE	SHANTELL'S CHANNEL - https://www		
10	JzCsM1vtn78	17.14.11	THE LOGANG MADE HISTORY. I	Logan Paul Vlogs	24	2017-11-1	logan paul	4477587	292837	4123	36391	https://i.yt	FALSE	FALSE	FALSE	Join the movement. Be a Maverick â		
11	43sm-QwLcx4	17.14.11	Finally Sheldon is winning an ar	Sheikh Musa	22	2017-11-1	God|"Shel	505161	4135	976	1484	https://i.yt	FALSE	FALSE	FALSE	Sheldon is roasting pastor of the chu		
12	H1KBHFXm2Bg	17.14.11	21 Savage - Bank Account (Offic	21 Savage	10	2017-11-1	21 savage	5068229	263596	8585	28976	https://i.yt	FALSE	FALSE	FALSE	Watch the official music video of Ba		
13	U3xLOo-CNwo	17.14.11	12 Weird Ways To Sneak Food I	Troom Troom	26	2017-11-1	sneak foo	3153224	28451	2285	3312	https://i.yt	FALSE	FALSE	FALSE	Subscribe Here: http://bit.ly/2uaz0or		
14	FyZMnhUtLfE	17.14.11	çŒŽåœª | Game Of Hunting 12	â¤§åŠ‡ç¨æ'-	1	2017-11-1	é›»è¦–å§	158815	218	30	186	https://i.yt	FALSE	FALSE	FALSE	Thanks for watching the drama! Help		
15	7MxiQ4v0EnE	17.14.11	Daang (Full Video) | Mankirt A	Speed Records	10	2017-11-1	punjabi so	5718766	127477	7134	8063	https://i.yt	FALSE	FALSE	FALSE	Song - Daang\nSinger - Mankirt Aulak		
16	LUzsOyWp9lw	17.14.11	YOUTUBERS REACT TO TOP 10	FBE	24	2017-11-1	twitter|"t	960747	31810	668	5335	https://i.yt	FALSE	FALSE	FALSE	CLICK TO SUBSCRIBE TO THE YOUTU		
17	AS9-ITLhQxo	17.14.11	I Hired An MI6 Spy To Help Me I	BuzzFeedBlue	22	2017-11-1	buzzfeed|'	1531218	53961	1697	4277	https://i.yt	FALSE	FALSE	FALSE	In the Outsmarted finale, Mike trains		
18	gjfPYwArCVQ	17.14.11	Fake Pet Smart Employee Prank	NELK	23	2017-11-1	prank|"pra	557883	44558	621	9619	https://i.yt	FALSE	FALSE	FALSE	3 Days left to cop NELK merch: https		
19	8NHA23f7LvU	17.14.11	Jason Momoa Wows Hugh Grat	The Graham Norton	24	2017-11-1	Graham N	1496225	16116	236	605	https://i.yt	FALSE	FALSE	FALSE	I think Sarah Millican was very excite		

LIKE RATIO PREDICTION METHODOLOGY



DERVING INSIGHTS FROM TEXT: METHODOLOGY

DATA PREPARATION

- Date Variables Transformation
- Publish Time

```
publish_time = pd.to_datetime(us_videos.publish_time, format='%Y-%m-%dT%H:%M:%S.%fZ')
```

- Trending Time

```
us_videos['trending_date'] = pd.to_datetime(us_videos['trending_date'], format='%y.%d.%m').dt.d
ate
```

- Add variable called “Trend_Time” that calculates difference between Trending Date & Publish Date

DATA PREPARATION

- Add Variable called “Likes_Ratio”

```
us_videos['like_ratio'] = us_videos['likes'] / (us_videos['dislikes'] + us_videos['likes'])
```

- Assign categorical numbers to “Categories”

```
1 - Film & Animation
2 - Autos & Vehicles
10 - Music
15 - Pets & Animals
17 - Sports
```

- Text Cleansing/Tokenization

TEXT PRE-PROCESSING EXAMPLE

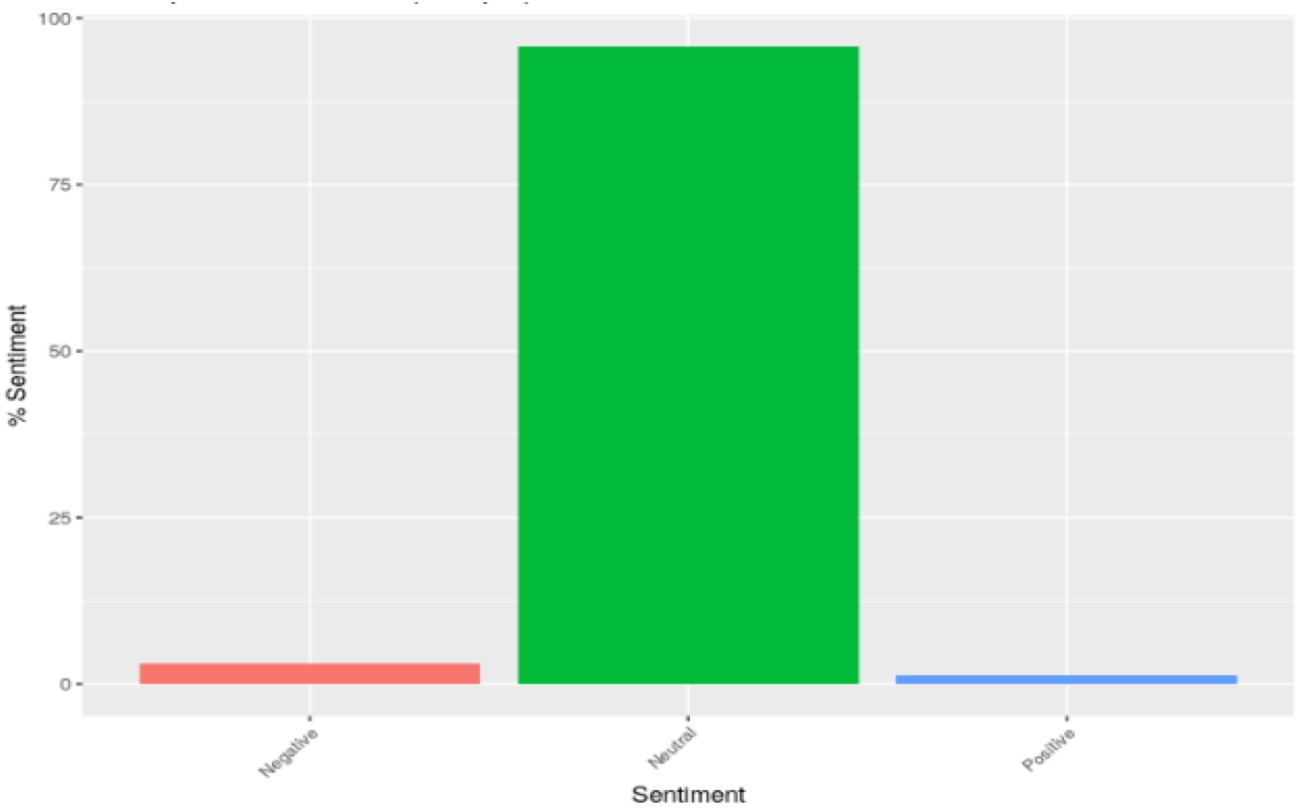
```
corpus2=[]
for i in range(0, 30000):

    review = re.sub('[^a-zA-Z]', ' ',dataset2['comment_text'][i] ) #kept all the letters from A to Z
    review = re.sub(r"[,@\'?\.$%_|]"," ",review,flags=re.I)
    review = review.lower()   # converted the letters to lower case
    review = review.split()
    ps = PorterStemmer()
    #review = [ps.stem(word) for word in review if not word in set(stopwords.words('english'))] #Stem function will extract the root
    #words out like 'love' from the word loved. But here it is not useful.
    review = [word for word in review if not word in set(stopwords.words('english'))]
    review = ' '.join(review)
    corpus2.append(review)
corpus2
```

```
['tad arrogant title huh',
 'apologize political correctness newspeak word leftist censorship media outlets trying bring solely seem somewhat right winged fucking wall street journal
 'pewdiepie bumass nigga lmao',
 'night face recognition huh',
 'highkey copied samsung',
 'best movie ever',
 'think ben carson needs neurosurgery',
 'dog u fall',
 'ohhh likes',
 'true many levels kinda hurts',
 'guys honest day school please like see',
 'accurate',
 'love comedy',
```

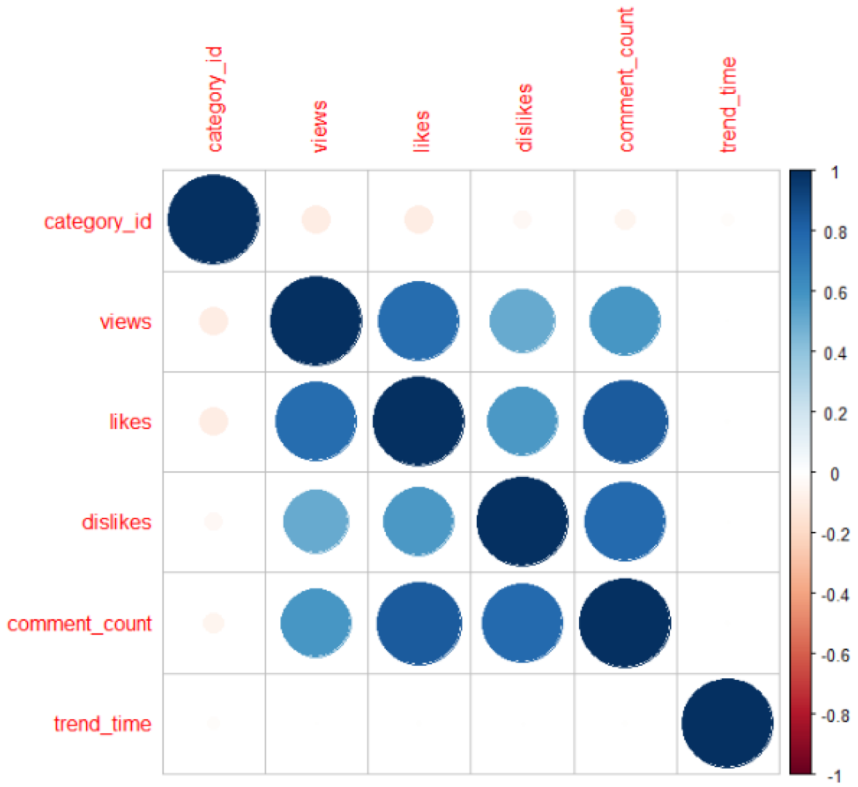
SENTIMENT ANALYSIS

- Also known as opinion mining
- Determines emotional tone behind a series of words
- General applications: social media monitoring, understanding public opinion

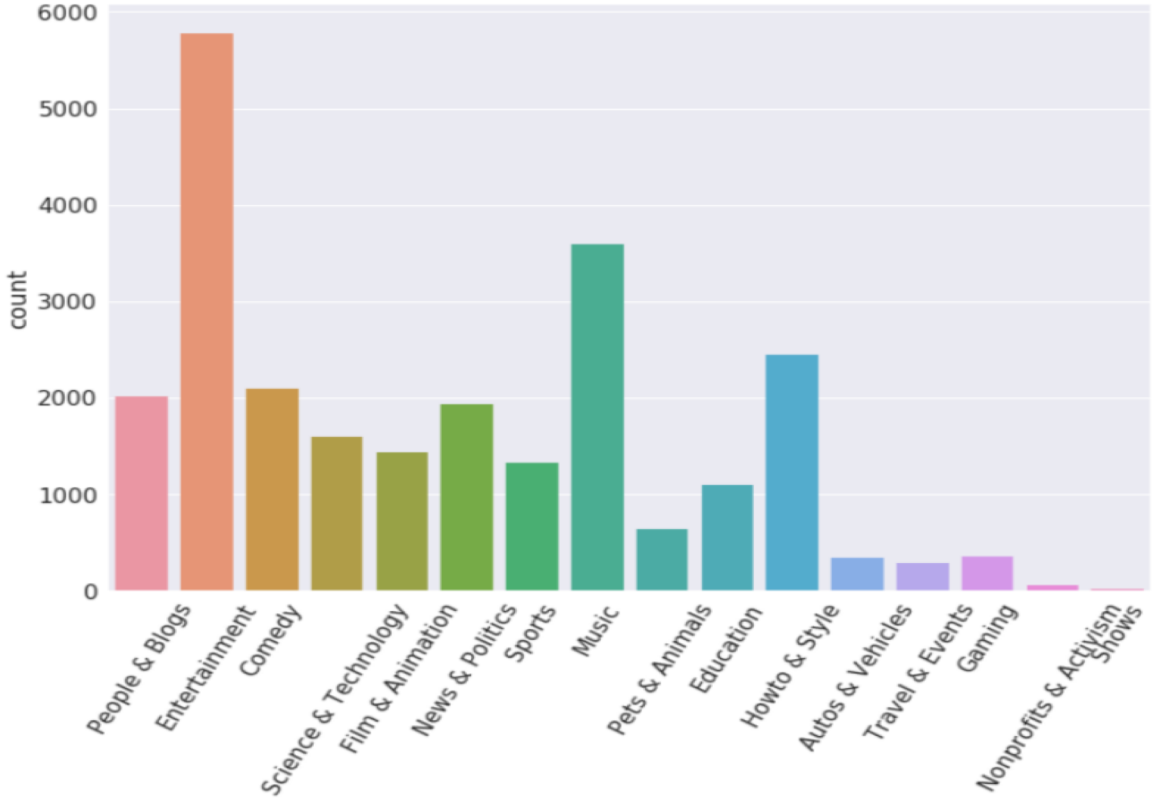
SENTIMENT ANALYSIS



EXPLORATORY ANALYSIS

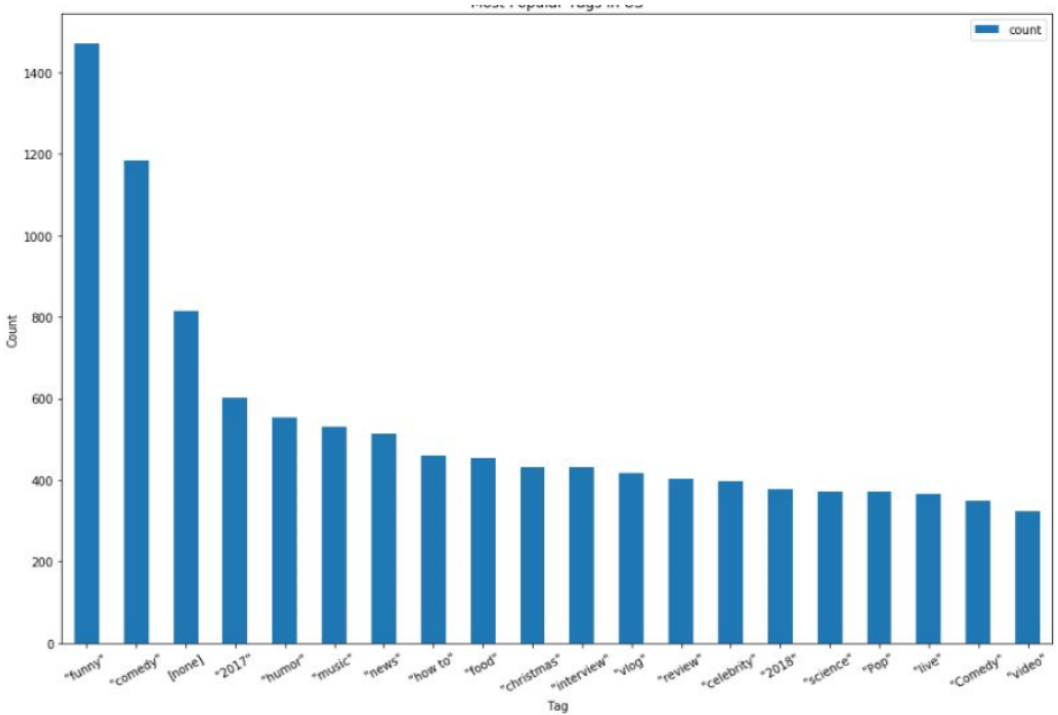


COUNT OF VIDEOS BY CATEGORIES

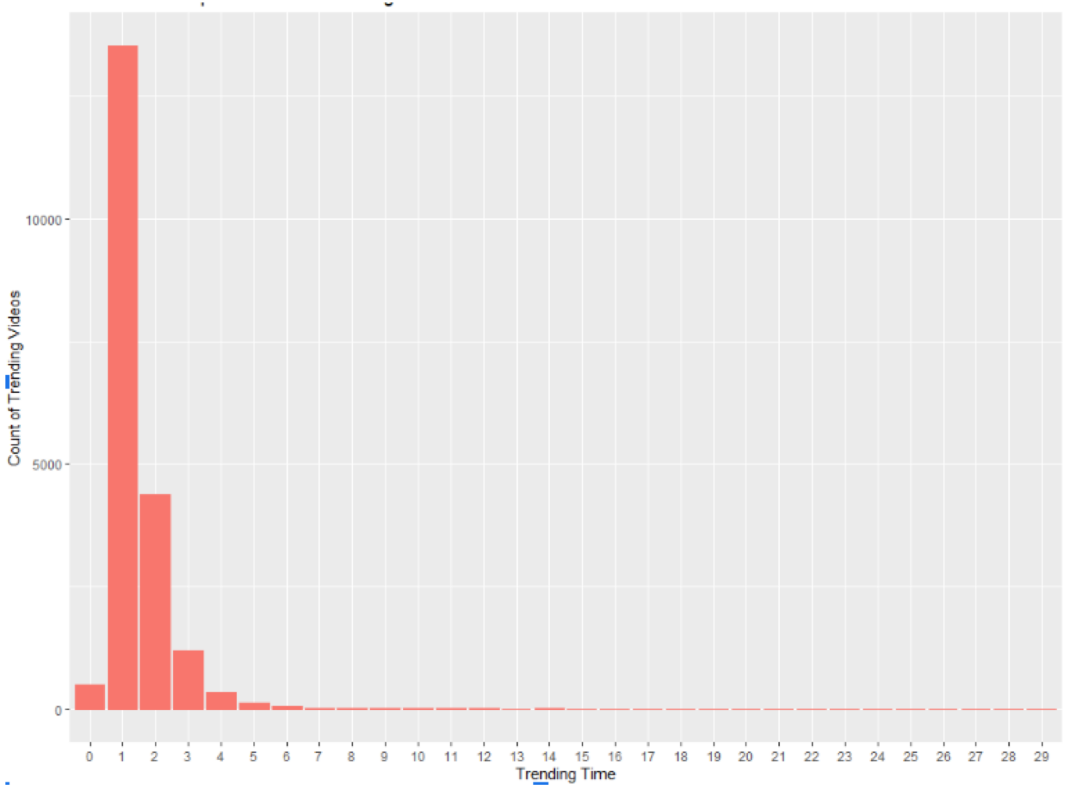


WORD CLOUD FOR POPULAR TAGS

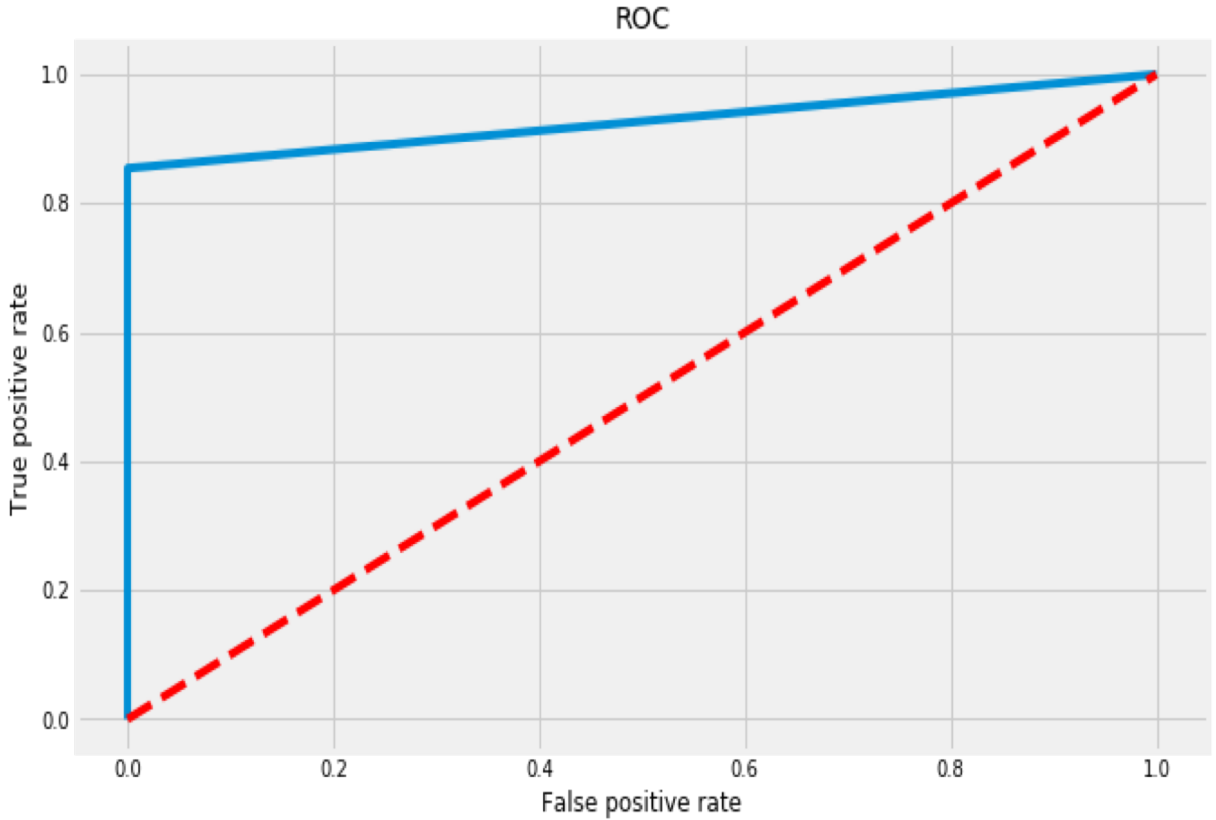
POPULAR TAG COUNT BY CATEGORY



COUNT OF TRENDING VIDEOS OVER TIME

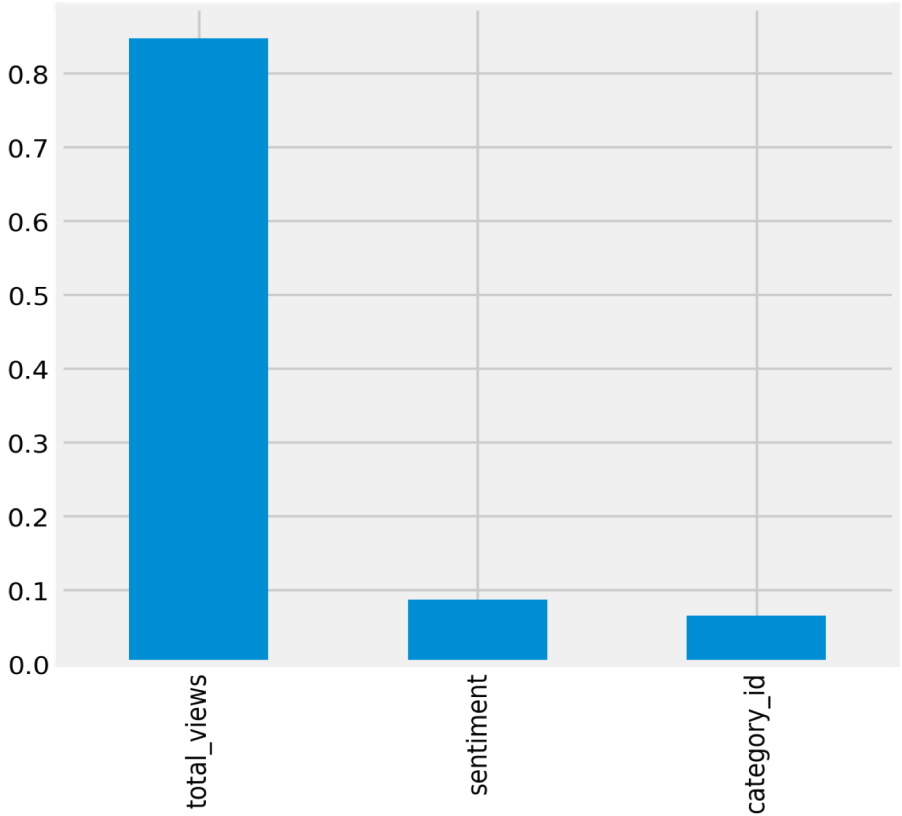


NAÏVE BAYES MODEL RESULTS



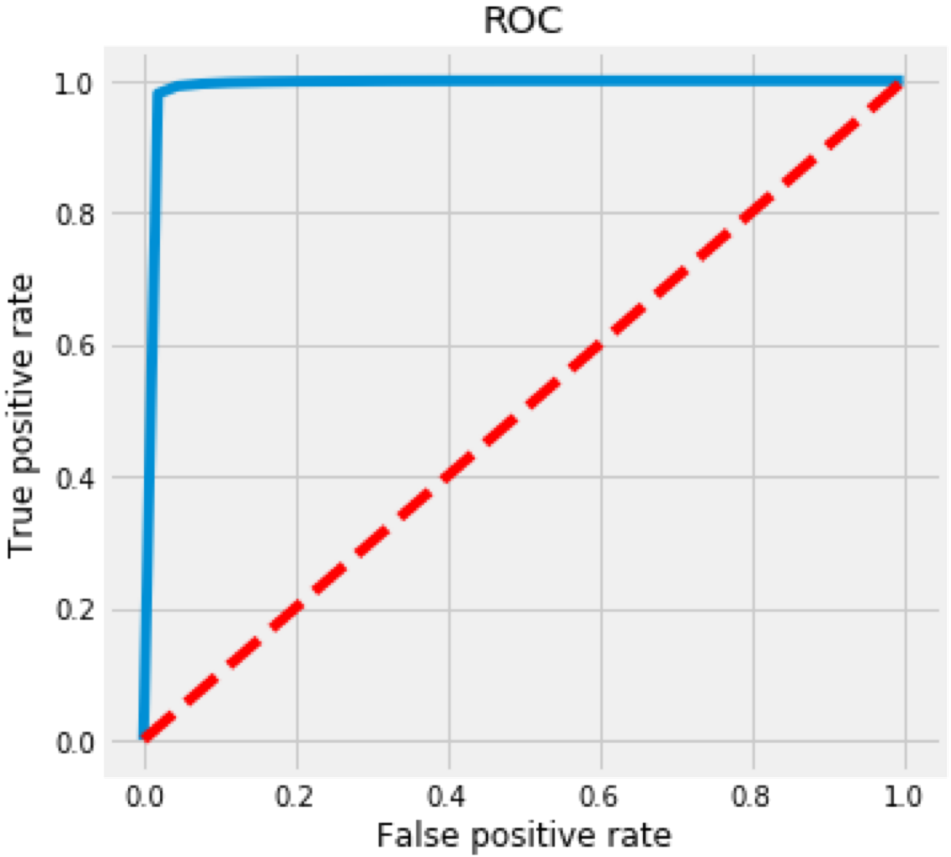
Precision: 0.85
Recall: 0.85
Fscore: 0.85
Accuracy:0.85
AUC: 0.927

RELATIVE IMPORTANCE OF VARIABLES



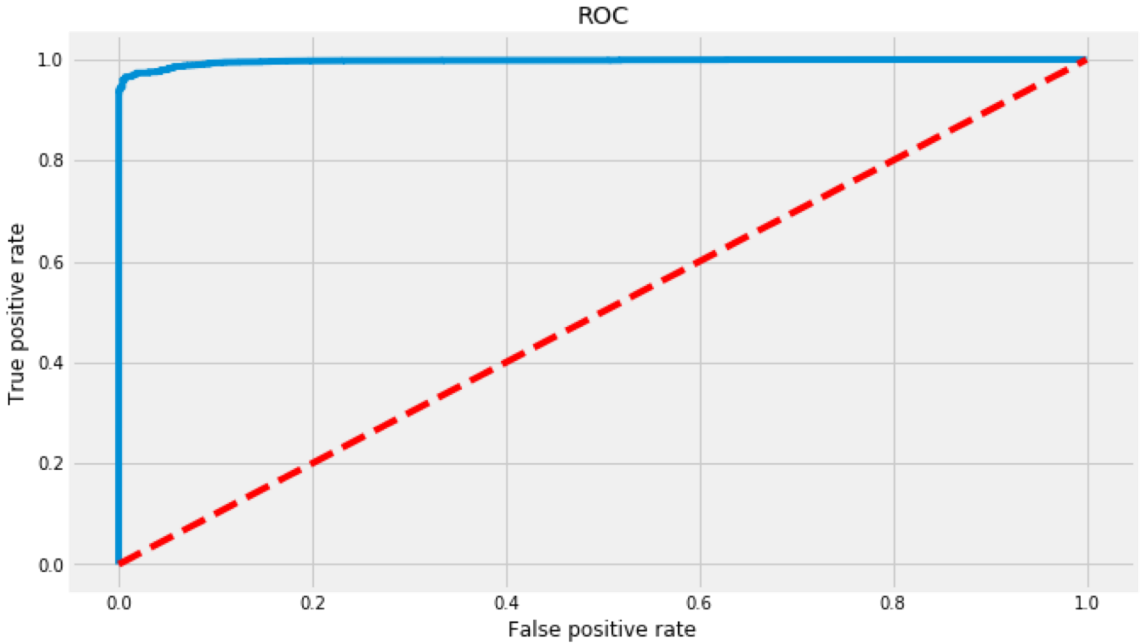
- Relative importance graph generated using ‘Random Forest’
- 'total_views’ appears as the most important predictor

RANDOM FOREST MODEL RESULTS



Precision: 0.98
Recall: 0.98
Fscore: 0.98
AUC: 0.98

LOGISTIC REGRESSION MODEL RESULTS



Precision: 0.98
Recall: 0.98
Fscore: 0.98
AUC: 0.739

RESULTS SUMMARY

- Maximum number of comments indicated neutral sentiments
- When the time gap between date of publishing and trending was 1 maximum number of videos were trending
- Most frequently used words were identified through generation of word cloud
- Random forest and Logistic Regression models used Views and sentiment as predictors
- The Random forest model predicted likeness ratio better than the logistic regression model by about 33%
- The Naïve Bayes Model helped use ‘comments’ to predict likeness ratio with about 93% efficiency

LIMITATIONS AND FUTURE SCOPE

Limitations:

1. Sentiment analysis is not 100% accurate and reliable and needs monitoring.
2. Human sentiments are not limited to positive, negative and neutral
3. Many of the NLP tools, packages or models are trained on social media data like twitter and Facebook and using the same sentiment classifiers for a unique dataset like YouTube may not give accurate results
4. Future scope involves using deep learning models to do sentiment analysis and likeness prediction

CONCLUSION

- Sentiment analysis was done on comments made by viewers and tested for ability to predict likeness ratio
- Shallow machine learning models were built
- Views, sentiment and comments were all found to be good predictors
- Further customization and improvement in the machine learning model can be obtained through using deep learning models

REFERENCES

https://pdfs.semanticscholar.org/1ad2/ef79142ac7ecb83aa8a2d9c77876ddbc95a0.pdf

https://www.kaggle.com/datasnaek/youtube-new