# Data Science Capstone Project

## YouTube Trending Video Analysis

Presented By:

Imama Mansoor

every
**MINUTE**
*of the* **DAY**

**Skype** USERS MAKE
**110,040** CALLS

**UBER** PASSENGERS TAKE RIDES **694**

**Facebook** USERS LIKE
**4,166,667** POSTS

**TWITTER** USERS SEND
**347,222** TWEETS

**Buzzfeed** USERS VIEW
**34,150** VIDEOS

**YOUTUBE** USERS UPLOAD
**300** HOURS OF NEW VIDEO

**SNAPCHAT** USERS SHARE
**284,722** SNAPS

**Instagram** USERS LIKE
**1,736,111** PHOTOS

**Tinder** USERS SWIPE
**590,278** TIMES

**PINTEREST** USERS PIN
**9,722** IMAGES

**VINE** USERS PLAY
**1,041,666** VIDEOS

**AMAZON** RECEIVES
**4,310** UNIQUE VISITORS

**reddit** USERS CAST
**18,327** VOTES

**Netflix** SUBSCRIBERS STREAM
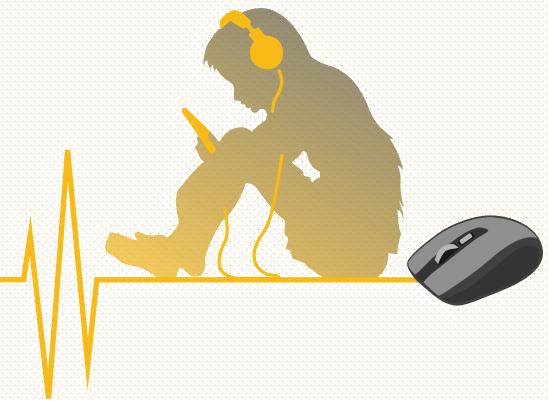**77,160** HOURS OF VIDEO

**APPLE** USERS DOWNLOAD
**51,000** APPS

# You Tube Trending Videos

"Trending helps viewers see what's happening on YouTube and in the world. Trending aims to surface videos that a wide range of viewers would find interesting. Some trends are predictable, like a new song from a popular artist or a new movie trailer. Others are surprising, like a viral video."

https://support.google.com/youtube/answer/7239739?hl=en

**The list of trending videos is updated roughly every 15 minutes.**

# Data Source and Tools used

You Tube Trending Video Data set from Kaggle having 16 features.

77,791 trending videos of United-States

From 11th Aug 2020 to 23rd August 2021, approx 200 videos per day.

Written on Google Colab using Python

# Project Pipeline

**Data Cleaning and Feature Engineering**

**Exploratory Data Analysis**

**Natural Language Processing**

**Likes Prediction using DistilBert and Keras Model**

# You Tube Trending Videos Dataset

Features
'video_id',
'title',
'publishedAt',
'channelId',
'channelTitle',
'categoryId',
'trending_date',
'tags',
'view_count',
'likes',
'dislikes',
'comment_count',
'thumbnail_link',
'comments_disabled',
'ratings_disabled',
'description'

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 77791 entries, 0 to 77790
Data columns (total 16 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   video_id           77791 non-null   object
 1   title              77791 non-null   object
 2   publishedAt        77791 non-null   object
 3   channelId          77791 non-null   object
 4   channelTitle       77791 non-null   object
 5   trending_date      77791 non-null   object
 6   tags               77791 non-null   object
 7   view_count         77791 non-null   int64
 8   likes              77791 non-null   int64
 9   dislikes           77791 non-null   int64
 10  comment_count      77791 non-null   int64
 11  thumbnail_link     77791 non-null   object
 12  comments_disabled  77791 non-null   bool
 13  ratings_disabled   77791 non-null   bool
 14  description        77791 non-null   object
 15  category           77791 non-null   object
dtypes: bool(2), int64(4), object(10)
memory usage: 9.1+ MB
```
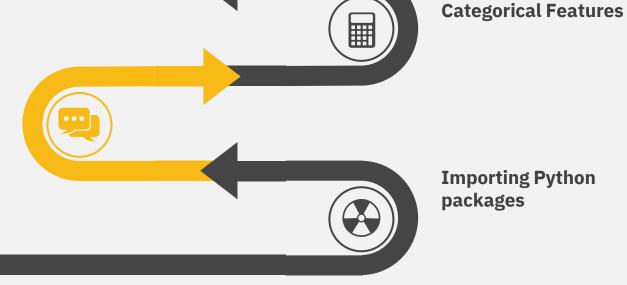
# Data Preparation

**Exploratory Data Analysis**
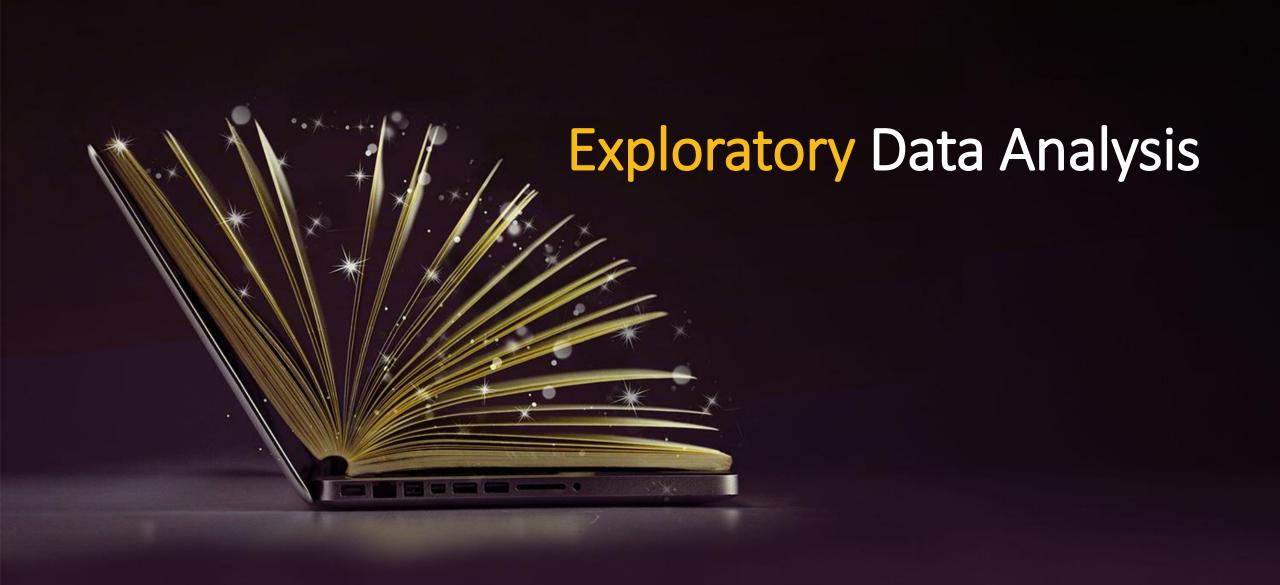
**Correlation between features**

**Distribution of Numerical and Categorical Features**

**Imputing missing and null values**
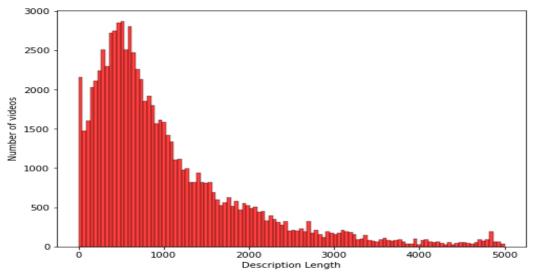
**Importing Python packages**

Exploratory Data Analysis
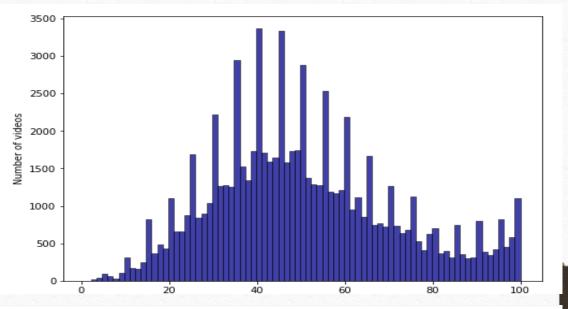
**Trending Videos' title lengths**
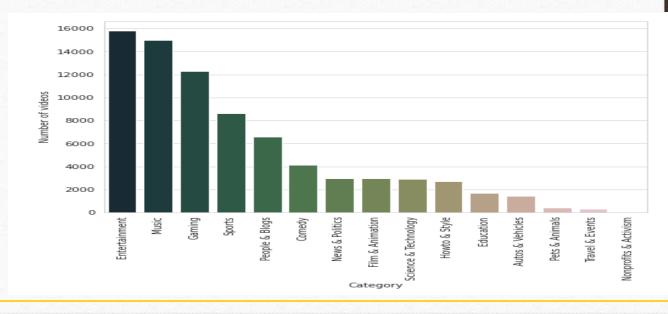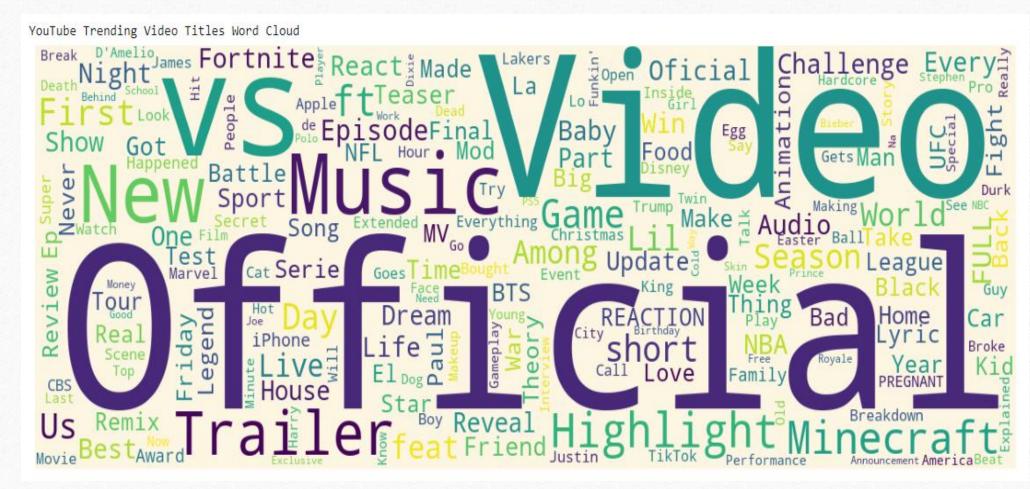
**Trending Videos' Description lengths**

# Trending for 36 Days!

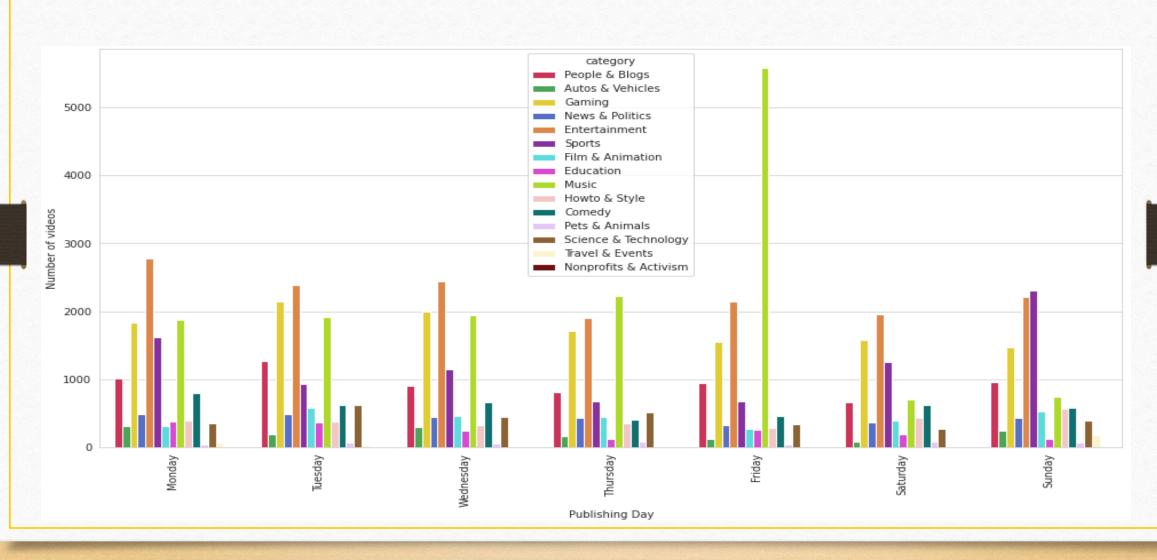## Which Video Category has the largest number of trending videos?

| | video_id | title | trending_days | view_count | likes | dislikes | comment_count | category |
|---|---|---|---|---|---|---|---|---|
| 66405 | hdmx71UjBXs | Turn into orbeez - Tutorial #Shorts | 36 | 4473783 | 190996 | 11442 | 120 | Entertainment |
| 59791 | aONT7atzqfw | Floyd Mayweather vs Logan Paul: Fight goes the... | 35 | 12255299 | 119173 | 16455 | 22494 | Sports |
| 31201 | wY6UyatwVTA | India claim stunning series win, end Australia... | 34 | 13516854 | 492839 | 8512 | 35888 | Sports |
| 53808 | mcUpeaSX8BE | Israeli Iron Dome filmed intercepting rockets ... | 34 | 1997878 | 27833 | 4333 | 0 | News & Politics |
| 67211 | vrRfBmMWpLY | Golden Buzzer: 9-Year-Old Victory Brinker Make... | 33 | 4271852 | 96708 | 2700 | 7159 | Entertainment |
| 7401 | pvPsJFRGleA | Justin Bieber - Holy ft. Chance The Rapper | 27 | 6217404 | 936304 | 13331 | 91366 | Music |
| 60392 | CZJvBfoHDk0 | Golden Buzzer: Nightbirde's Original Song Make... | 24 | 4021571 | 153813 | 1249 | 13865 | Entertainment |
| 64605 | 0sgk1sNUJR8 | Golden Buzzer: Jimmie Herrod Surprises the Jud... | 23 | 3765394 | 96602 | 1742 | 6841 | Entertainment |

# Word Cloud



YouTube Trending Video Titles Word Cloud
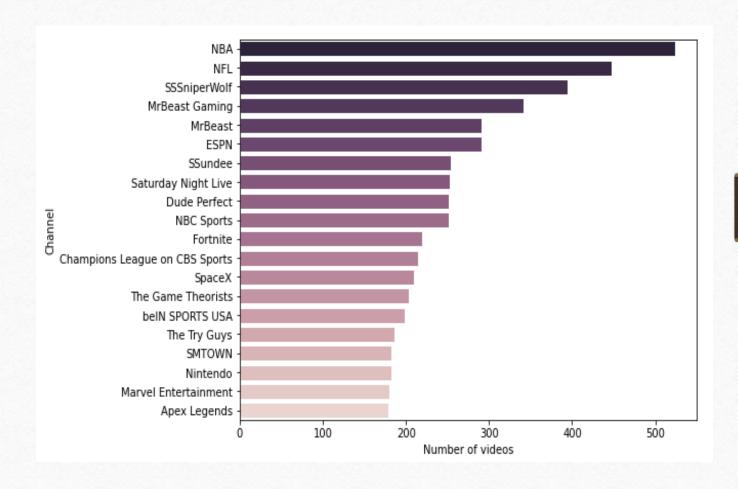
# Most Popular Publishing Day as per Category
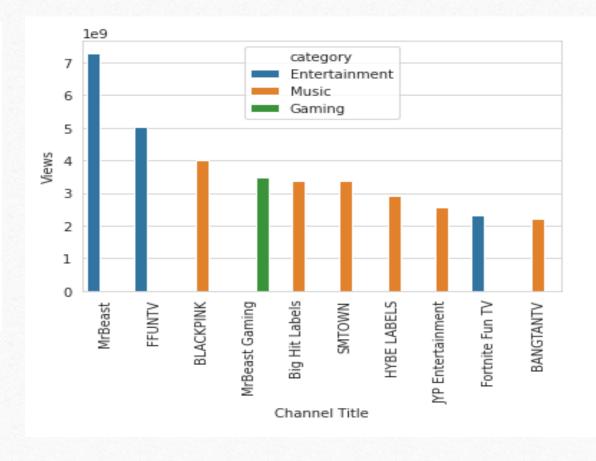
# Which Channels have largest number of trending videos?

| | channelTitle | video_count |
|------|---|---|
| 2277 | NBA | 524 |
| 2298 | NFL | 447 |
| 2811 | SSSniperWolf | 395 |
| 2244 | MrBeast Gaming | 342 |
| 2243 | MrBeast | 291 |
| 979 | ESPN | 291 |
| 2812 | SSundee | 254 |
| 2856 | Saturday Night Live | 253 |
| 951 | Dude Perfect | 252 |
| 2285 | NBC Sports | 251 |
| 1183 | Fortnite | 220 |
| 594 | Champions League on CBS Sports | 215 |
| 2988 | SpaceX | 210 |
| 3225 | The Game Theorists | 203 |
| 3734 | beIN SPORTS USA | 198 |
| 3298 | The Try Guys | 186 |
| 2802 | SMTOWN | 182 |
| 2392 | Nintendo | 182 |
| 2110 | Marvel Entertainment | 180 |
| 203 | Apex Legends | 179 |
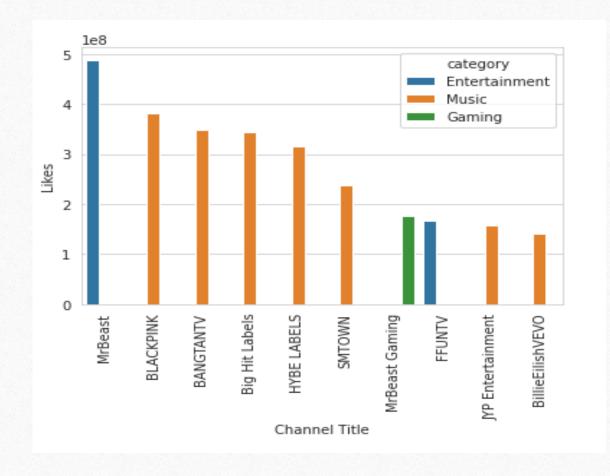
# Channels with the most View count



|   | channelTitle | category | view_count |
|---|--------------|----------|------------|
| 0 | MrBeast | Entertainment | 7285763030 |
| 1 | FFUNTV | Entertainment | 5030654339 |
| 2 | BLACKPINK | Music | 4010600470 |
| 3 | MrBeast Gaming | Gaming | 3484682421 |
| 4 | Big Hit Labels | Music | 3378335085 |
| 5 | SMTOWN | Music | 3361022584 |
| 6 | HYBE LABELS | Music | 2906191985 |
| 7 | JYP Entertainment | Music | 2555470483 |
| 8 | Fortnite Fun TV | Entertainment | 2328240867 |
| 9 | BANGTANTV | Music | 2210181691 |

# Most Liked channel on their Trending videos by Category
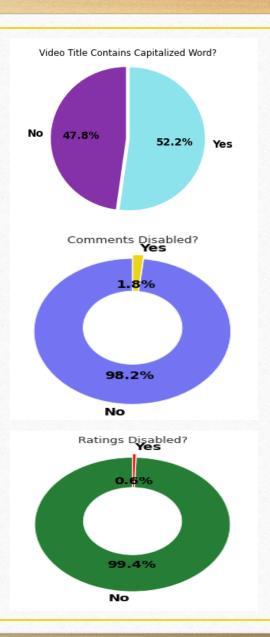


|   | channelTitle | category | likes |
|---|---|---|---|
| 0 | MrBeast | Entertainment | 488594961 |
| 1 | BLACKPINK | Music | 381873493 |
| 2 | BANGTANTV | Music | 349686698 |
| 3 | Big Hit Labels | Music | 343992023 |
| 4 | HYBE LABELS | Music | 316431776 |
| 5 | SMTOWN | Music | 237242992 |
| 6 | MrBeast Gaming | Gaming | 177641694 |
| 7 | FFUNTV | Entertainment | 167738799 |
| 8 | JYP Entertainment | Music | 157810404 |
| 9 | BillieEilishVEVO | Music | 142179689 |

**How many trending video titles contain capitalized word?**
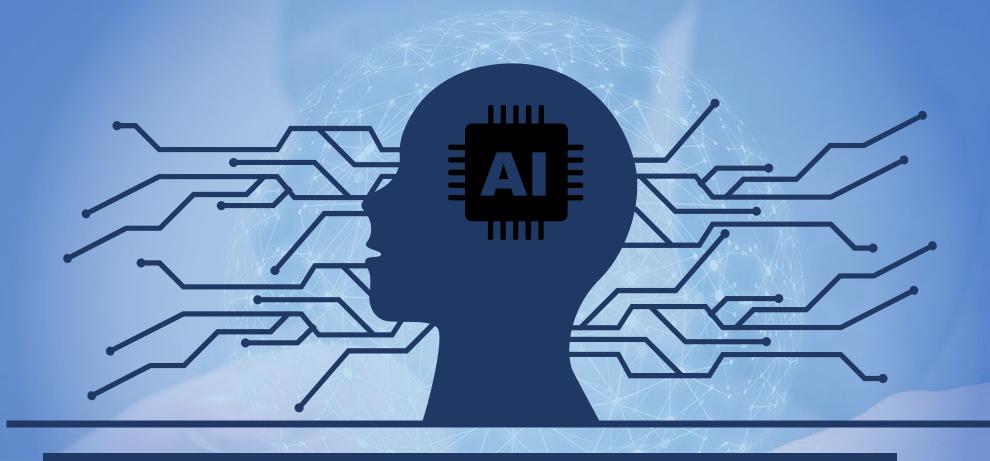
**How many Trending videos have their comments disabled?**
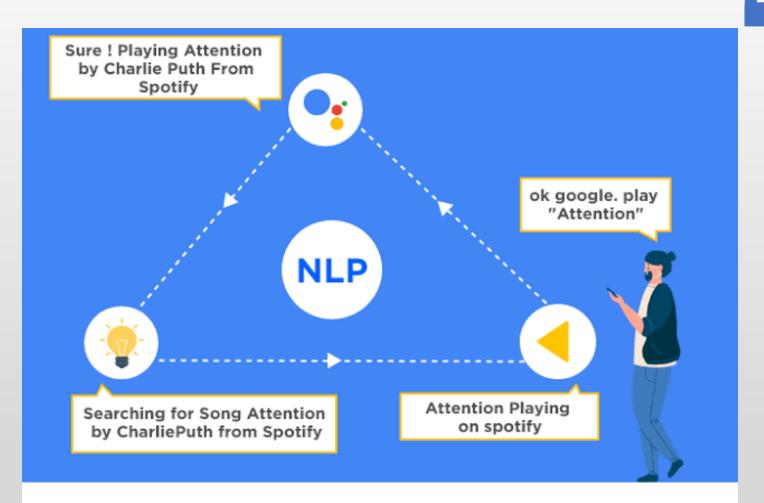
**How many Trending videos have their ratings disabled?**

Video Title Contains Capitalized Word?

No 47.8%    52.2% Yes

Comments Disabled?
Yes
1.8%
98.2%
No

Ratings Disabled?
Yes
0.6%
99.4%
No

# Some more interesting Insights:

- Trending top most video with maximum view count is **'Butter' Official MV.**
- **IPhone 12 and IPhone 12 Pro!** is on trending with the least views.
- **'Butter Official MV'** is the most liked video having **16021534** likes.
- **BLACKPINK - Icecream ( with Selena Gomez) MV** is the most disliked trending video having **879354** dislikes.
- Number of trending videos published on weekends i.e Sunday and Saturday are noticeably less than the number of trending videos published on other days of the week.
- **552** videos were published on **02 June 2021** which got on trending!!!
- The time period between 2PM and 10PM, peaking between 3PM and 5PM, had the largest number of trending videos.
- Mostly videos on trending requires 3 days to get in the trending list, but some videos may require from 12 to 35 days to get viral and be in trending!
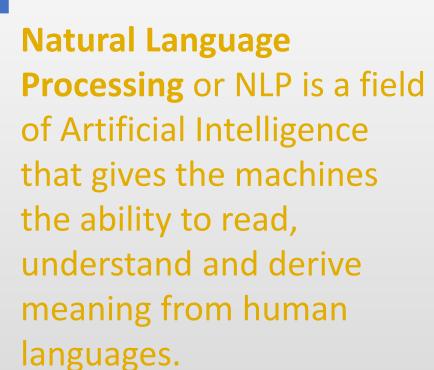
# Natural Language Processing
( NLTK Package Python)

**Natural Language Processing** or NLP is a field of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human languages.
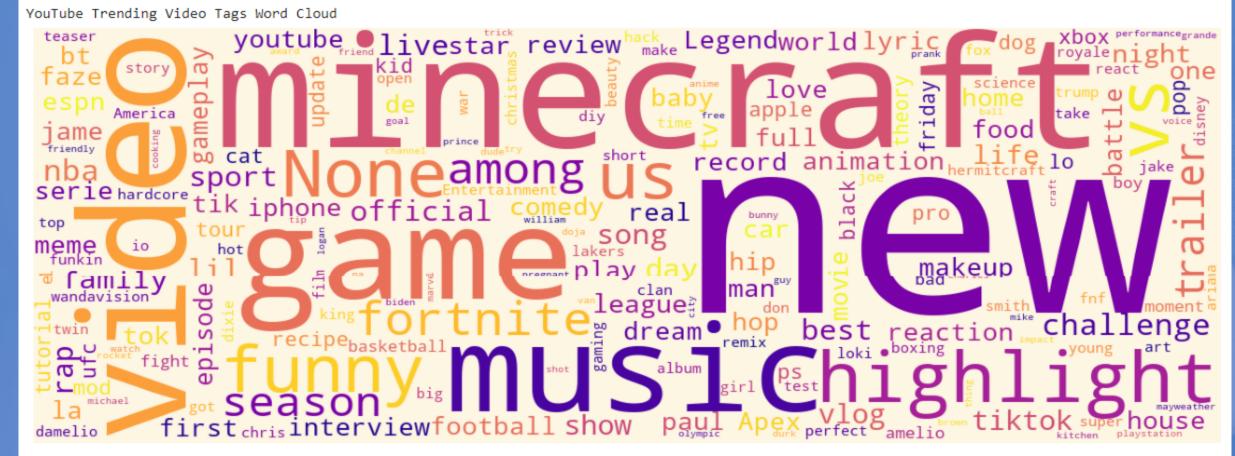
# **Tokenization**

[('the', 24304),
 ('minecraft', 17191),
 ('new', 13030),
 ('and', 13016),
 ('video', 12277),
 ('to', 11950),
 ('vs', 11906),
 ('None', 11157),
 ('highlights', 10190),
 ('funny', 10148)]

YouTube Trending Video Tags Word Cloud

# Bigrams

[(('(official', 'video)'), 3126),
(('(official', 'music'), 2144),
(('music', 'video)'), 2143),
(('highlights', '|'), 2000),
(('among', 'us'), 1221),
(('official', 'trailer'), 1019),
(('in', 'the'), 915),
(('(video', 'oficial)'), 843),
(('of', 'the'), 802),
(('music', 'video]'), 763)]

YouTube Trending Video Title Bigrams Word Cloud

# Applying Stop Words technique (NLTK) of NLP

[('official', 10625),
('video', 9149),
('vs', 4258),
('music', 3546),
('new', 2508),
('highlights', 2443),
('ft', 2441),
('minecraft', 2100),
('x', 1944),
('trailer', 1880)]

YouTube Trending Video Titles Exclusive Word Cloud

# Applying POS Tagging (NLTK) of NLP

NN ➔ noun, singular (cat, tree)
NNS ➔ noun plural (desks)
PRP ➔ personal pronoun (hers, herself, him, himself)
RB ➔ adverb (occasionally, swiftly)
VBD ➔ verb past tense (pleaded)
JJ ➔ adjective (large)…

```
[('official', 'NN')]
[('video', 'NN')]
[('vs', 'NN')]
[('music', 'NN')]
[('new', 'JJ')]
[('highlights', 'NNS')]
[('ft', 'NN')]
[('minecraft', 'NN')]
[('x', 'NN')]
[('trailer', 'NN')]
[('game', 'NN')]
[('first', 'RB')]
[('season', 'NN')]
[('feat', 'NN')]
[('us', 'PRP')]
[('live', 'JJ')]
[('lil', 'NN')]
[('among', 'IN')]
[('world', 'NN')]
[('day', 'NN')]
[('full', 'JJ')]
[('episode', 'NN')]
[('audio', 'NN')]
[('night', 'NN')]
[('oficial', 'JJ')]
[('baby', 'NN')]
[('nba', 'NN')]
[('life', 'NN')]
[('fortnite', 'NN')]
[('every', 'DT')]
[('part', 'NN')]
[('got', 'VBD')]
[('challenge', 'NN')]
[('best', 'JJS')]
```

# Top Most Emojis Used in Trending Video Titles



```
[('😀', 87),
 ('😂', 74),
 ('😱', 67),
 ('🔴', 67),
 ('😭', 37),
 ('◉', 37),
 ('❌', 34),
 ('👑', 34),
 ('🔥', 31),
 ('\u200d', 31)]
```

# DistilBERT Transformer Model

- Distilled version of BERT base model
- Smaller and faster than BERT
- Pretrained on the same corpus as BERT in a self-supervised fashion



## GOAL

✓ **Likes Prediction with the help of Text Classification by using Huggingface Transformers library**

# Custom Dataset for Likes Prediction



|   | description | likes |
|---|---|---|
| 0 | SUBSCRIBE to BRAWADIS ▶ http://bit.ly/Subscrib... | 156908 |
| 127 | As the generation that grew up playing Road Ra... | 30254 |
| 128 | We have a giant area of land and whatever you ... | 1643982 |
| 129 | After negotiations with Democrats in Congress ... | 6045 |
| 130 | Clan Wars 2 is coming. Get ahead of the compet... | 73921 |

|   | description | likesCat |
|---|---|---|
| 0 | subscribe to brawadis me on social twitter i... | 1 |
| 127 | as the generation that grew up playing road ra... | 0 |
| 128 | we have a giant area of land and whatever you ... | 1 |
| 129 | after negotiations with democrats in congress ... | 0 |
| 130 | clan wars 2 is coming get ahead of the competi... | 0 |
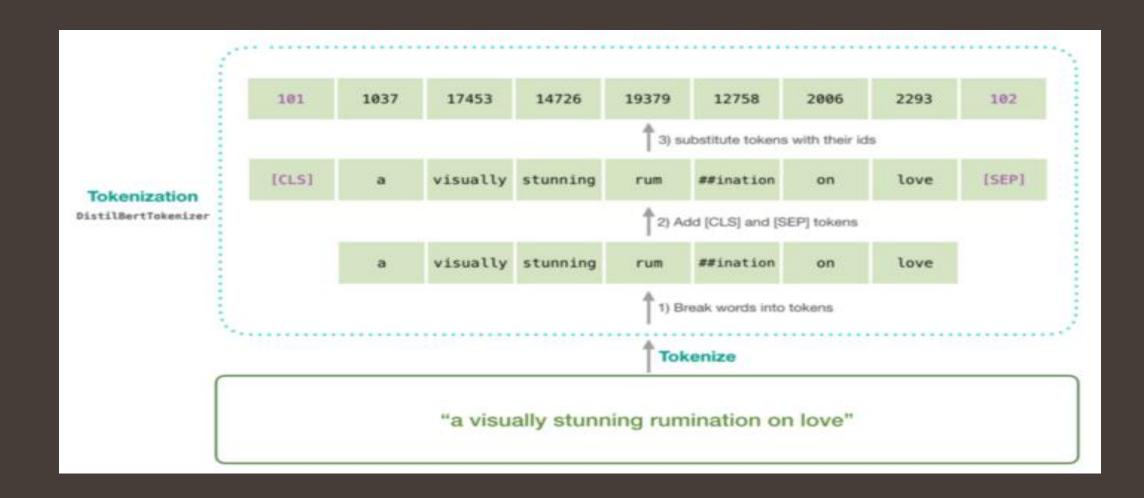
## LIMITATIONS

- Selected features
- Taking a sample of 1000 videos
- Convert to classification problem
  - 0 - Low for Likes <= 0.1 Million
  - 1 - High for Likes > 0.1 Million

# Distil Bert Tokenizer

# Model Pipeline

✓ Text Preprocessing using TextHammer package

✓ Splitting Train and Test Data from Dataset

✓ DistilBERT Tokenizer converts the Train data into input ids and attention masks

✓ Creating Word Embeddings through DistilBERT

✓ Implementing Keras Layers

✓ Instantiating Keras Model

✓ Model Compilation

✓ Model Fitting on our input ids and attention masks

✓ Model Prediction

```
{'input_ids': <tf.Tensor: shape=(700, 512), dtype=int32, numpy=
array([[  101,  4470,  5074, ...,      0,     0,     0],
       [  101,  3050, 20098, ...,      0,     0,     0],
       [  101, 15624,   102, ...,      0,     0,     0],
       ...,
       [  101,  1045,  7714, ...,      0,     0,     0],
       [  101,  2603, 29100, ...,      0,     0,     0],
       [  101,  2131,  2115, ...,      0,     0,     0]], dtype=int32)>, 'attention_mask': <tf.Tensor: shape=(700, 512), dtype=int32, numpy=
array([[1, 1, 1, ..., 0, 0, 0],
       [1, 1, 1, ..., 0, 0, 0],
       [1, 1, 1, ..., 0, 0, 0],
       ...,
       [1, 1, 1, ..., 0, 0, 0],
       [1, 1, 1, ..., 0, 0, 0],
       [1, 1, 1, ..., 0, 0, 0]], dtype=int32)>}
```

```
{'balanced_accuracy': [0.7685714364051819, 0.9200000166893005],
 'loss': [0.4888262152671814, 0.2133064190311432],
 'val_balanced_accuracy': [0.8766666650772095, 0.9233333468437195],
 'val_loss': [0.3216328024864197, 0.22794079780578613]}
```

# Model Evaluation

## Classification Report

|  | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.96 | 0.95 | 212 |
| 1 | 0.90 | 0.83 | 0.86 | 88 |
| | | | | |
| accuracy | | | 0.92 | 300 |
| macro avg | 0.92 | 0.90 | 0.91 | 300 |
| weighted avg | 0.92 | 0.92 | 0.92 | 300 |

Root Mean Square Error between True Labels and Predicted Labels for Test Data:
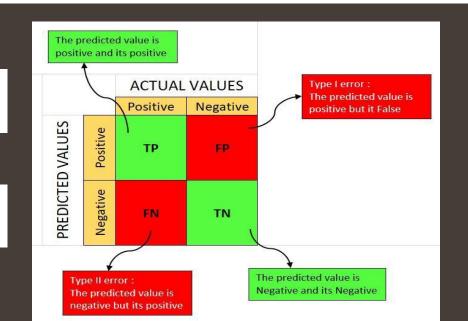
0.27688746209726917

## Confusion Matrix

Train Data

```
array([[490,    3],
       [ 18, 189]])
```

Test Data

```
array([[204,    8],
       [ 15,   73]])
```

# ITS TIME FOR 'LIKES' PREDICTION

Write the description of your video →

We have a giant area of land and whatever you build, I'll pay for it. SUBSCRIBE OR YOU'LL HAVE GOOD LUCK for the future  New Merch - shopmrbeast.com. Join our discord server for your chance to compete in future Minecraft challenges! You will definitely win!!!

Model Prediction

```
Low 28.385809
High 45.73733
```

**28% =>> Likes below 0.1 Million**
**45% =>> Likes above 0.1 Million**

# Challenges faced

❑ Limited resources for deep learning

❑ Haven't done NLP before

❑ Complex topics

❑ Tough debugging

# Learning Outcomes

o I can work on similar projects

o I'll try to take hands-on experience on other top NLP language models

LEARNING

# THANKS FOR YOUR PRECIOUS TIME

Presented By:

IMAMA MANSOOR