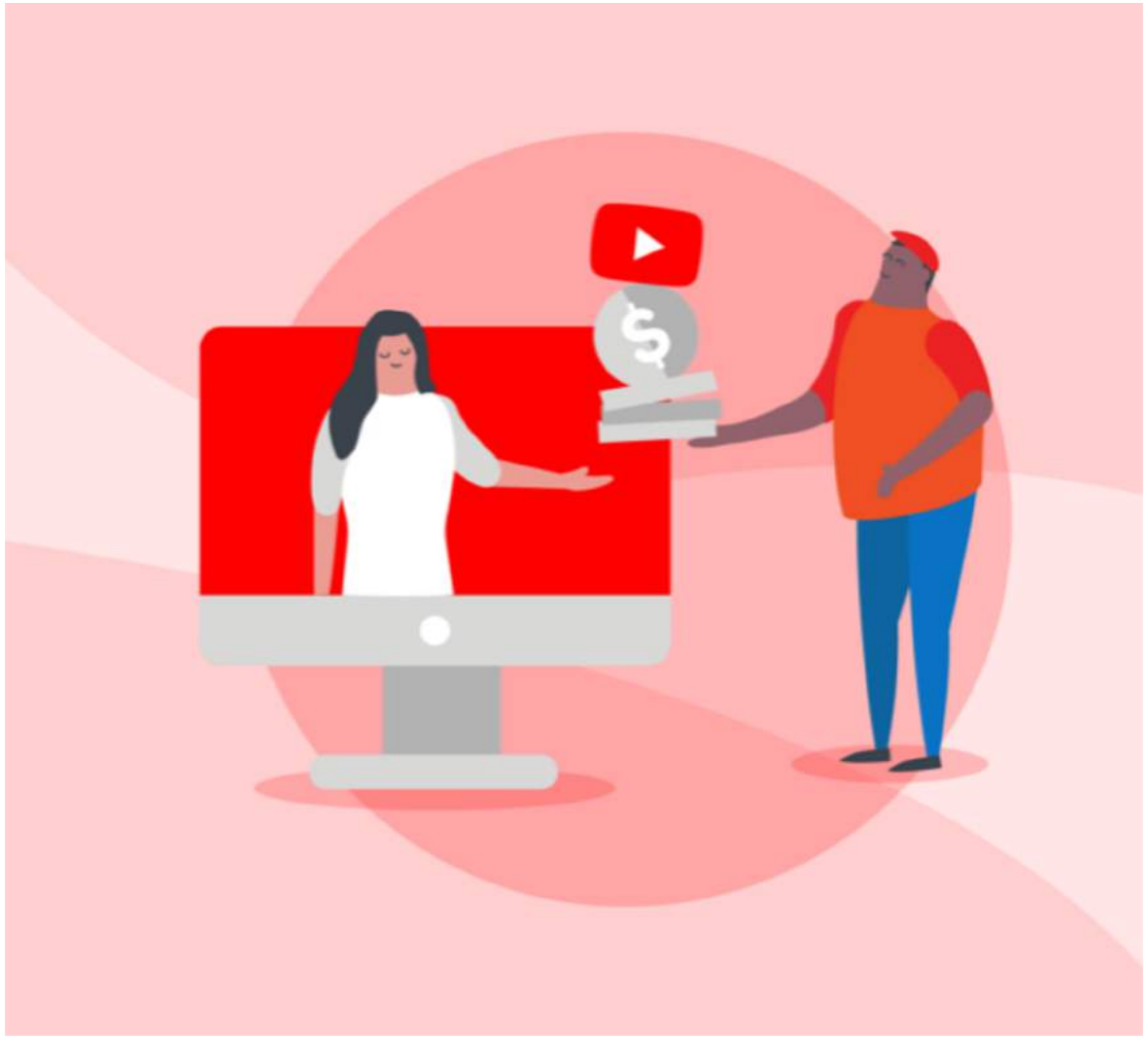
Analysis and Prediction of YouTube Trending Videos

by Ariel Li, Jiaying Du, Sylvie Pan, Xinyuan Gao

Agenda

- Business Problem
- Data Overview
- Exploratory Data Analysis
- Feature Engineering and Visualization
- Modeling and Prediction
- Future Work and Deployment



Business Overview

Business Case Summary

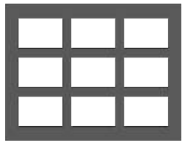
YouTube is the world's most popular video-sharing platform. Its popularity has prompted companies to put their ads on the website. Since the views accurately reflect the scope of audience an ad can reach, it is in the marketing companies' interest to find out whether a video will go viral to decide where to cast their campaigns on this video.

To achieve this goal, we analyzed the YouTube trending videos in 3 English-speaking countries, Canada, the UK and the US, conducted feature engineering and built a classification model to predict if a video will go viral.

Data Overview

Data Summary

Data Source:
Kaggle YouTube Trending Video Statistics



Datasets:
Canada (40881 records)
UK (38916 records)
US (40949 records)

Date Range:
November 2017 - June 2018

Number of Variables:
17

A Closer Look in Data

9 Features of Type Object:

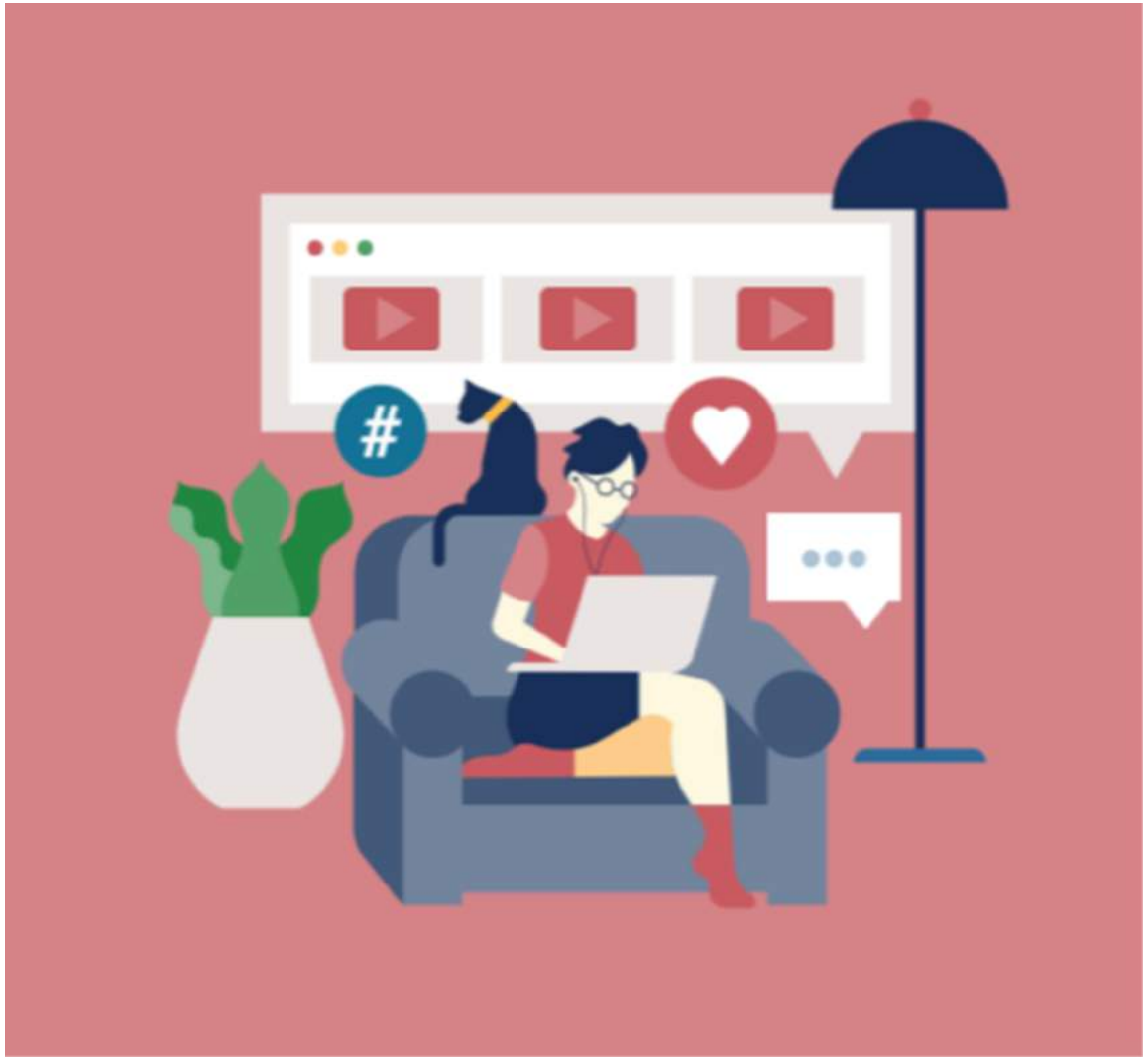
- String: Video ID, title, channel title, tags, thumbnail link, description, category
- Datetime: trending date, publish time

5 Features of Type Integer:

- Views, likes, dislikes, comment_count, category ID

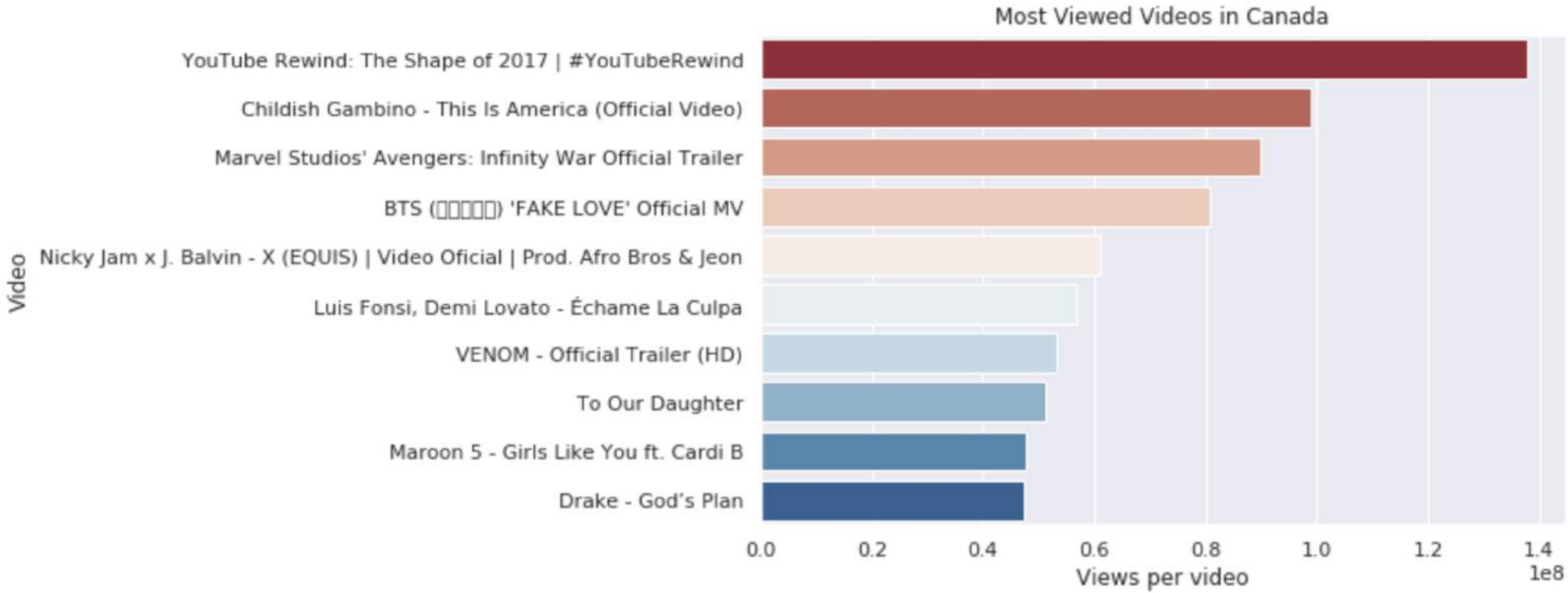
3 Features of Type Boolean:

- Comments_disabled, ratings_disabled, video_error_or_removed

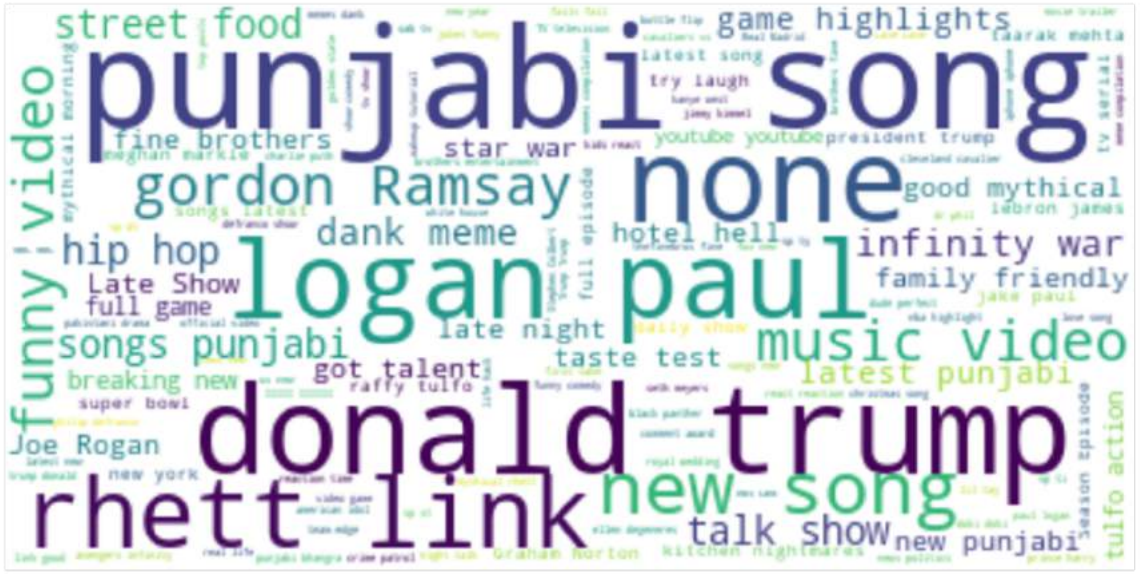


Exploratory Data Analysis

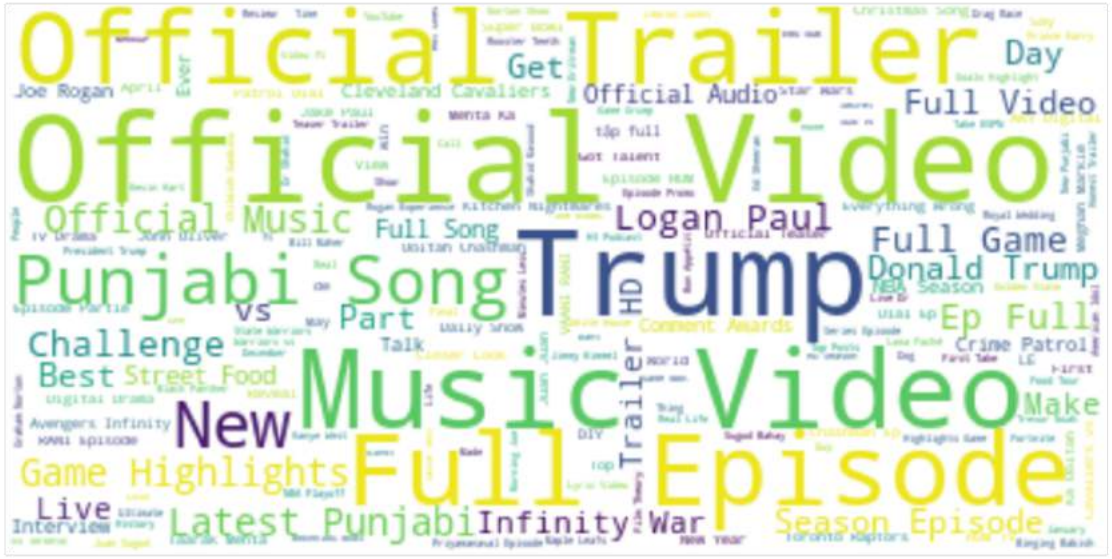
Most Viewed Videos in Canada



Word Clouds for Videos in Canada

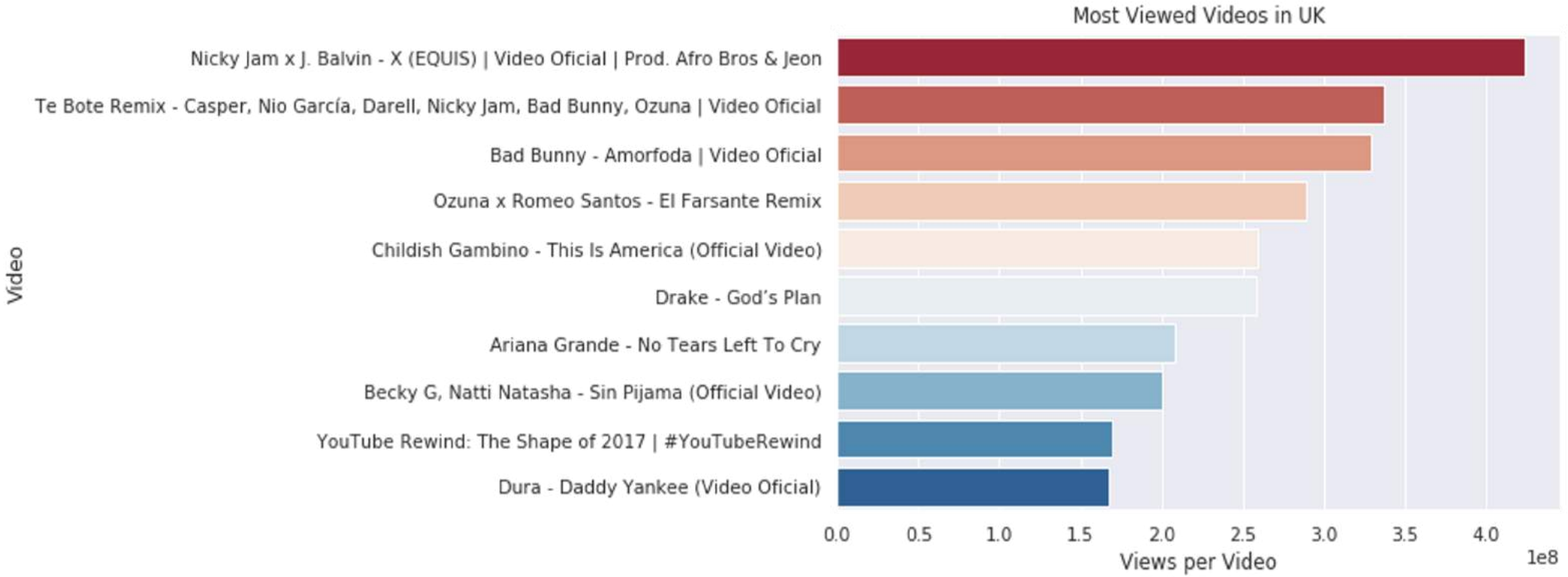


Word cloud for tags



Word cloud for titles

Most Viewed Videos in UK

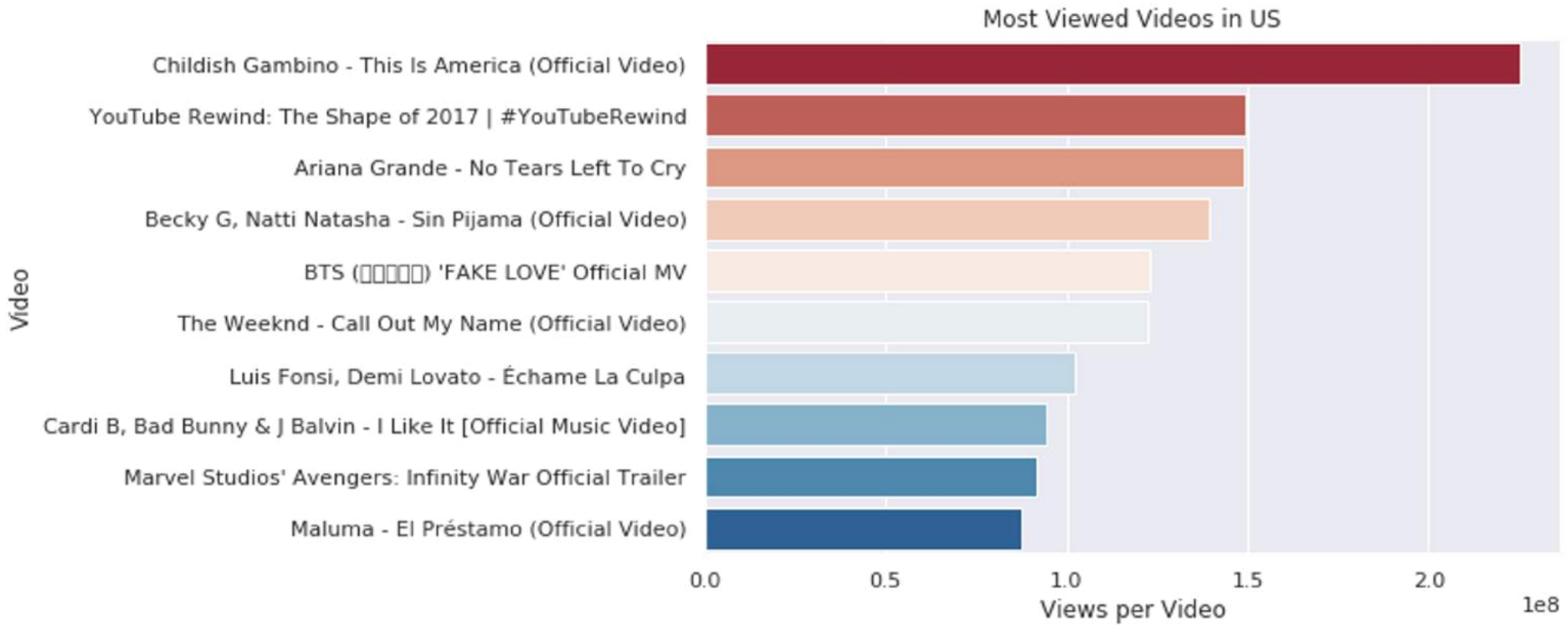


Word Clouds for Videos in UK

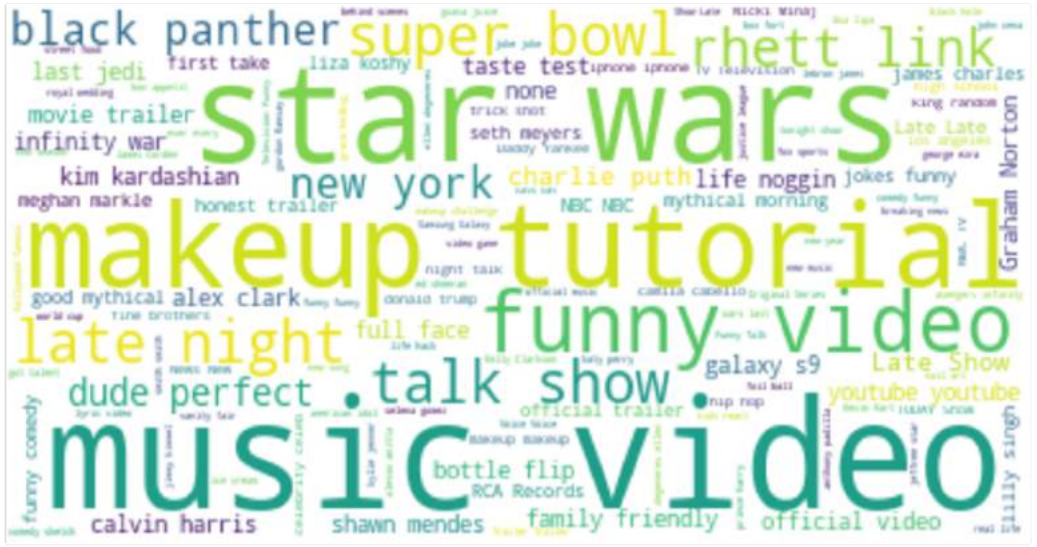
Word cloud for tags

Word cloud for titles

Most Viewed Videos in US



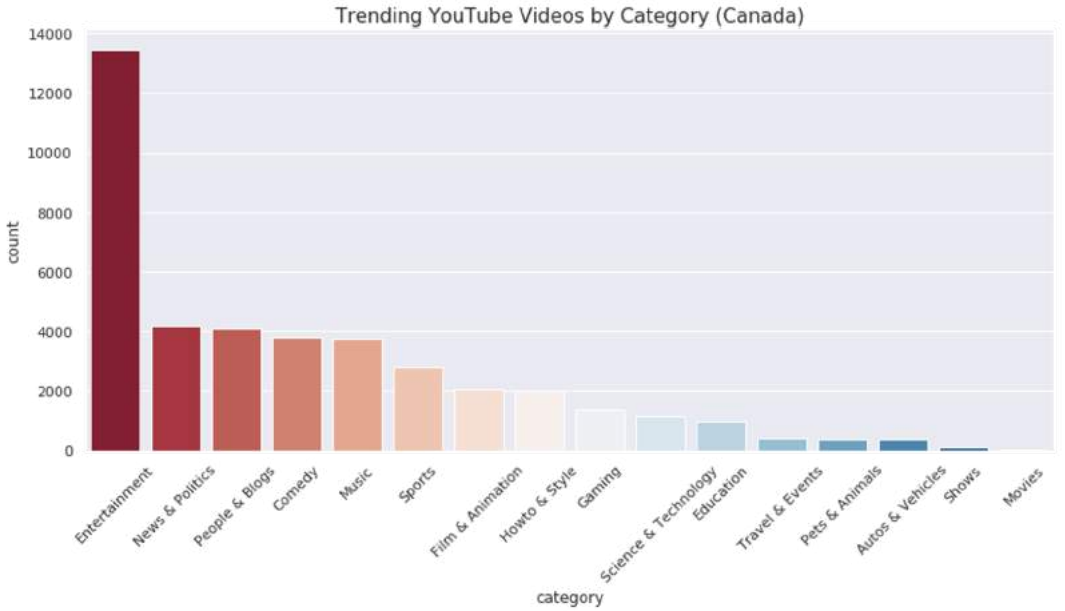
Word Clouds for Videos in US

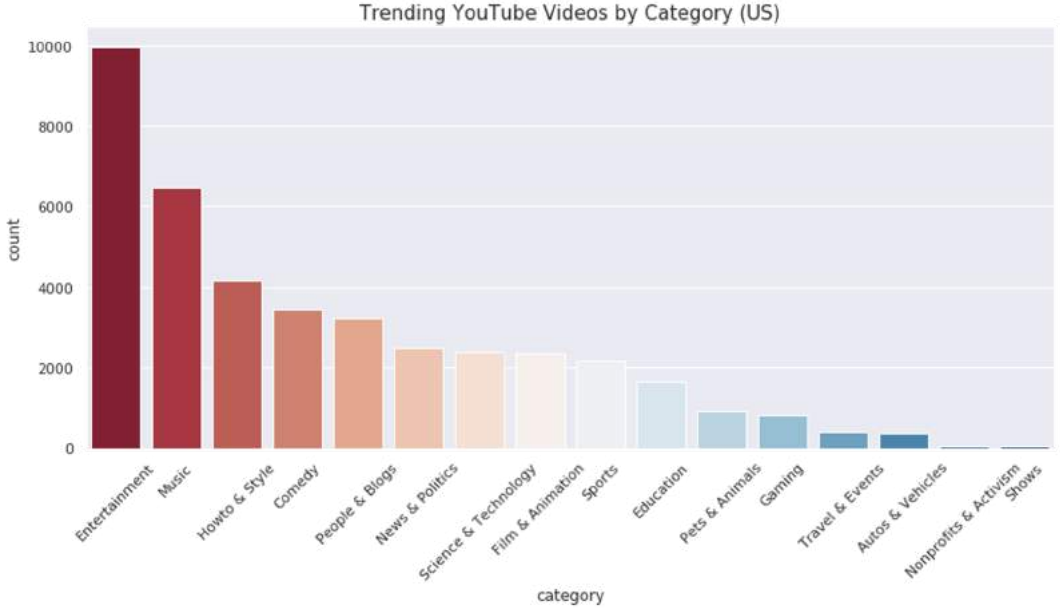


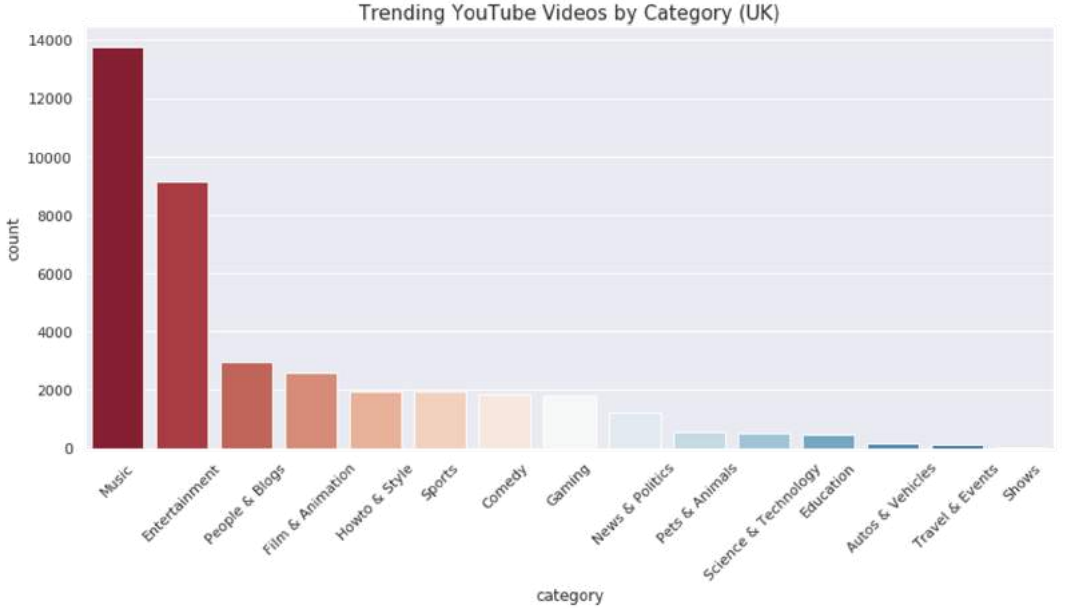
Word cloud for tags

Word cloud for titles

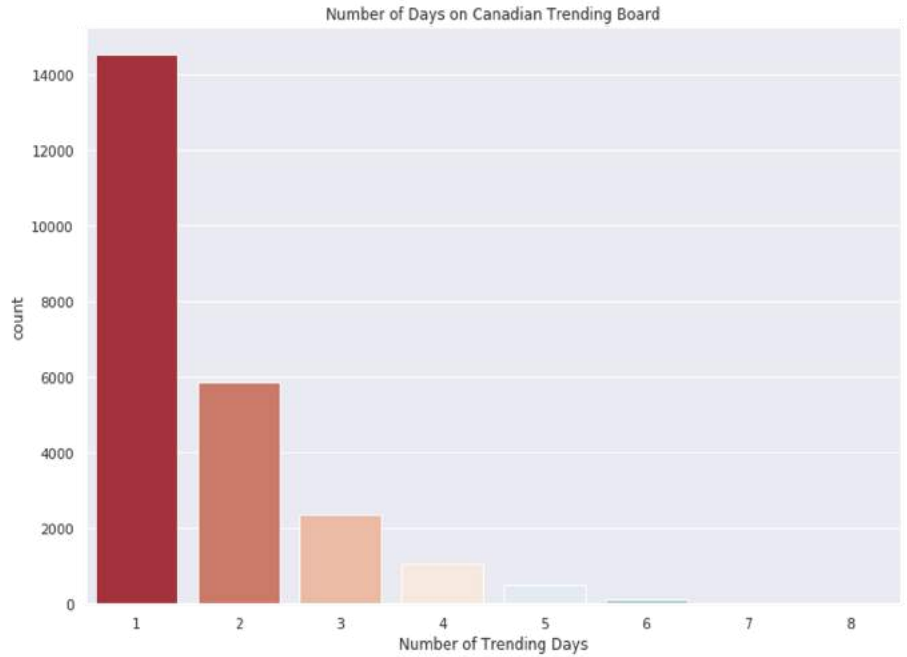
Categories of Trending Videos

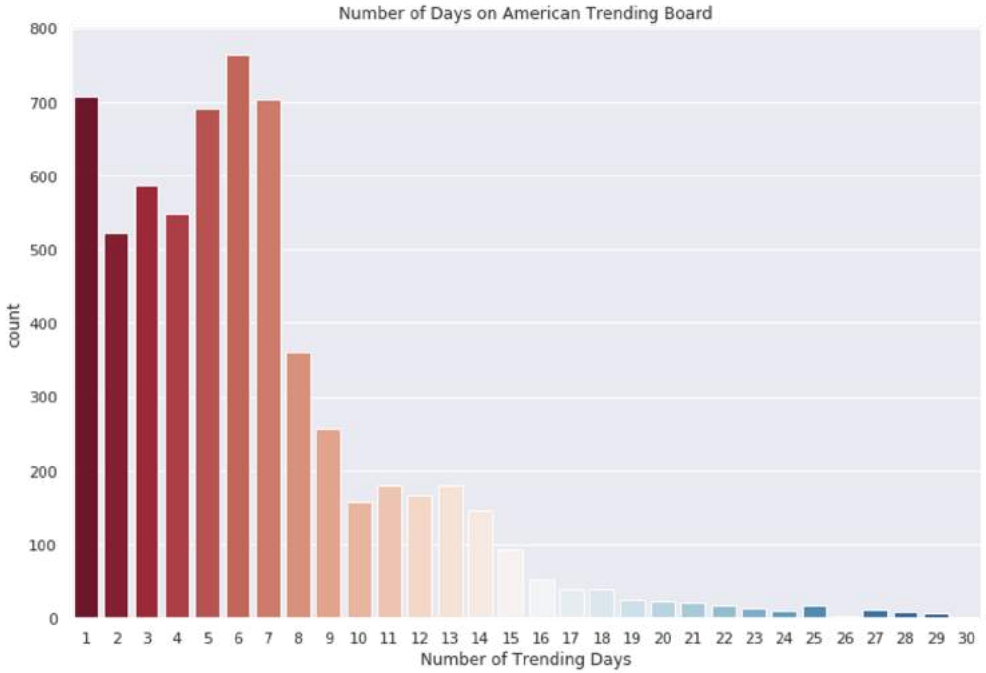


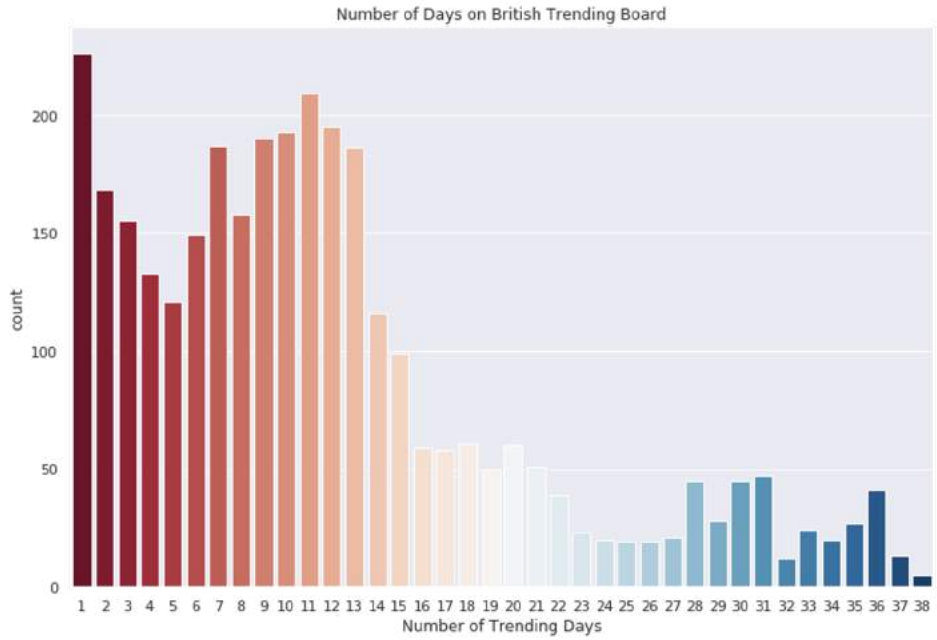




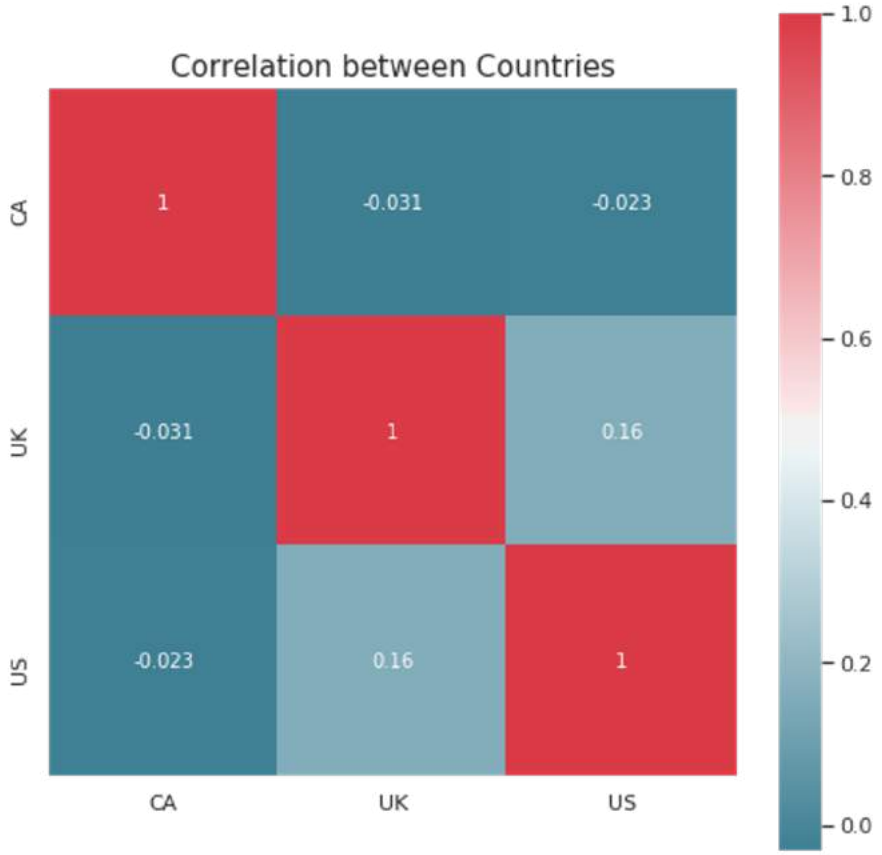
Number of Trending Days







Trending Videos Across Countries



Top Videos with Longest Combined Trending Duration

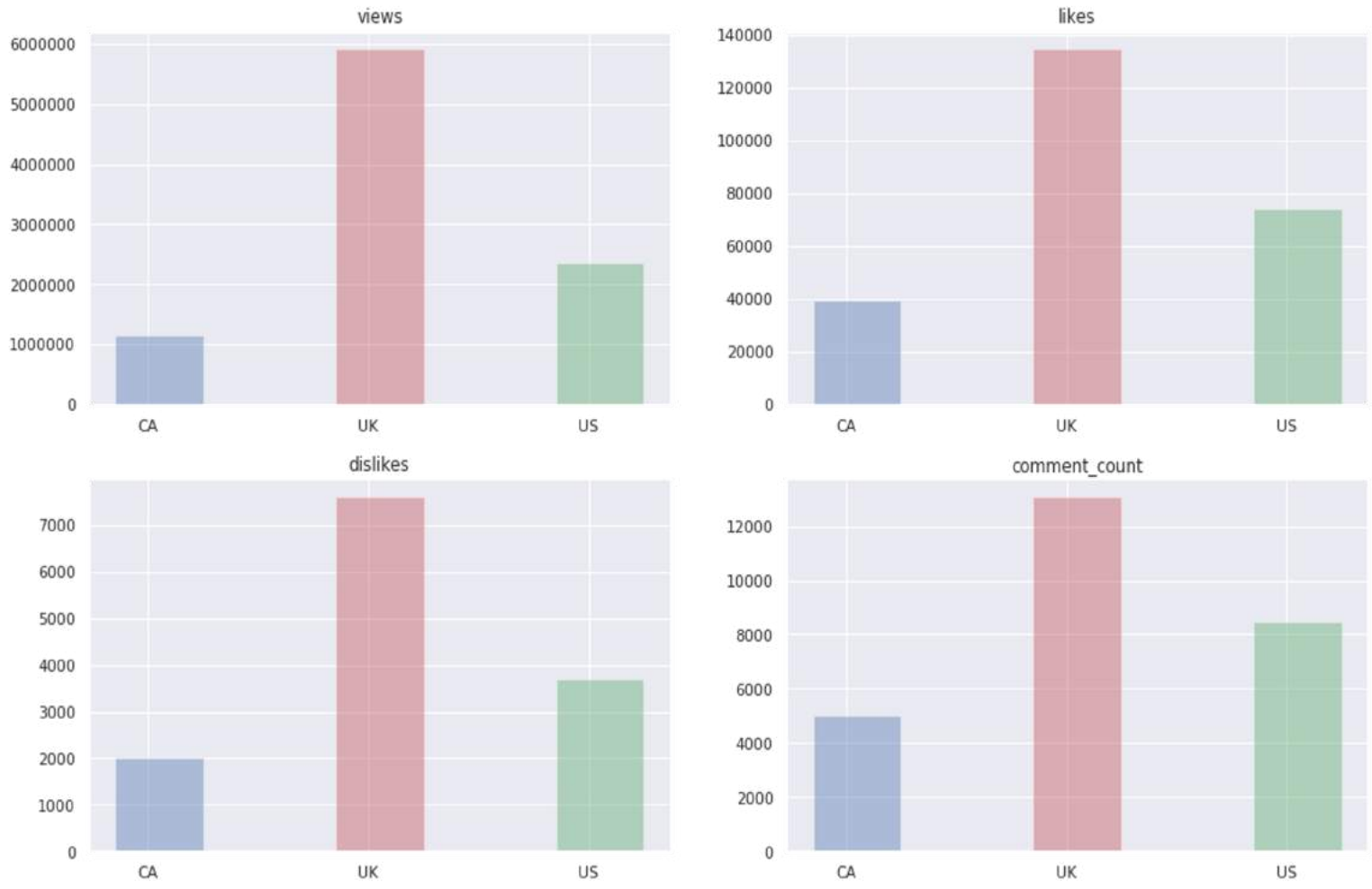
title	CA	UK	US	Total
Sam Smith - Pray (Official Video) ft. Logic	4.0	37.0	29.0	70.0
Childish Gambino - This Is America (Official Video)	8.0	36.0	25.0	69.0
Maroon 5 - Wait	9.0	40.0	18.0	67.0
Getting some air, Atlas?	3.0	34.0	28.0	65.0
Marvel Studios' Ant-Man and The Wasp - Official Trailer	5.0	36.0	23.0	64.0

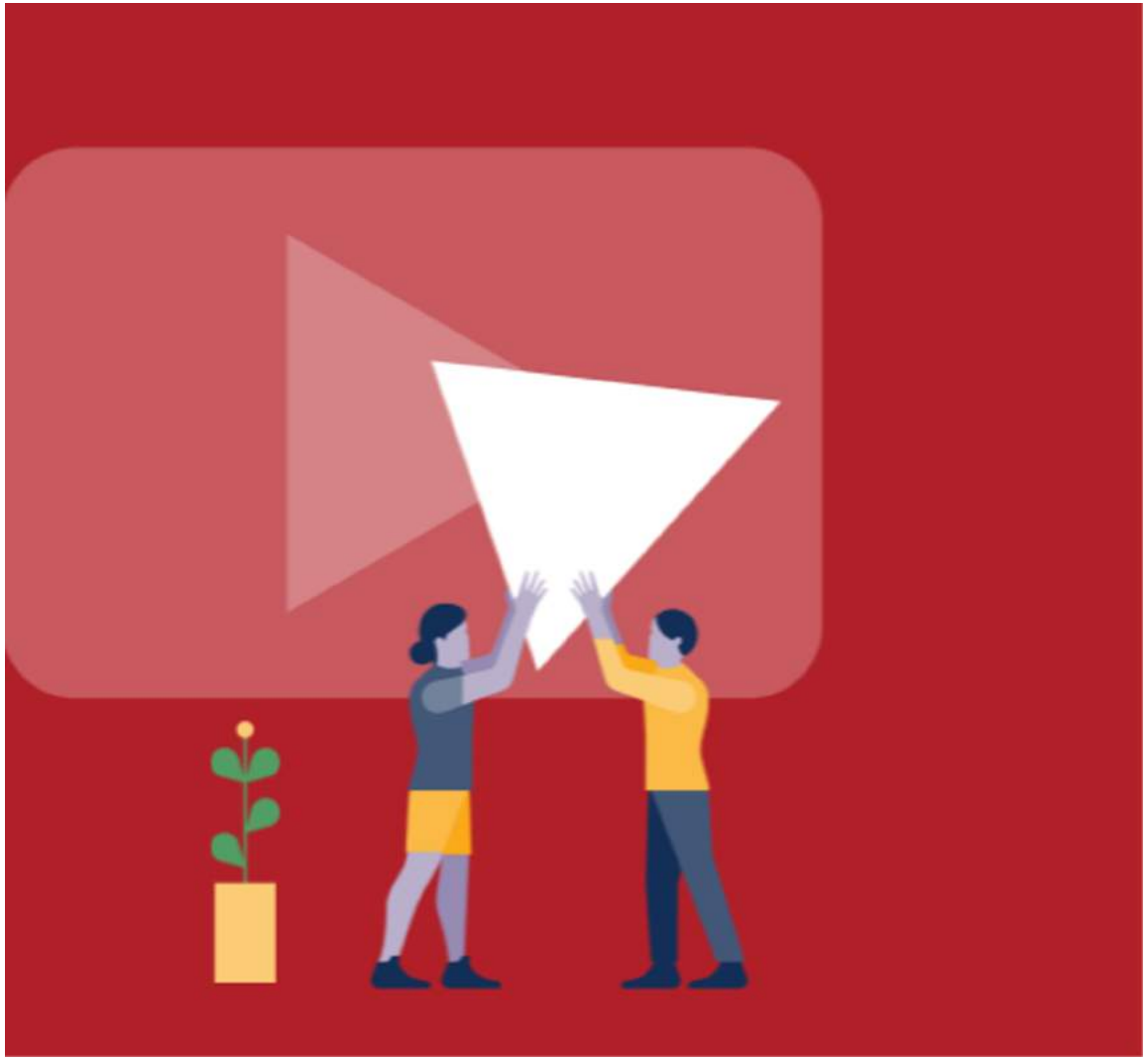
Videos Trended in >1 Country:

89%

Statistics by Country

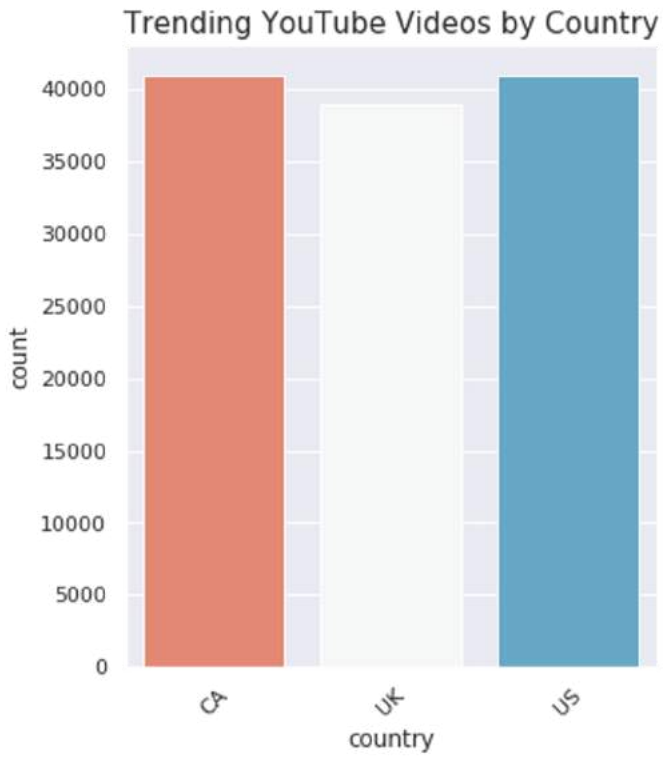
Average Views, Likes, Dislikes & Comment Counts





Feature Engineering and Visualizations

Merging Datasets



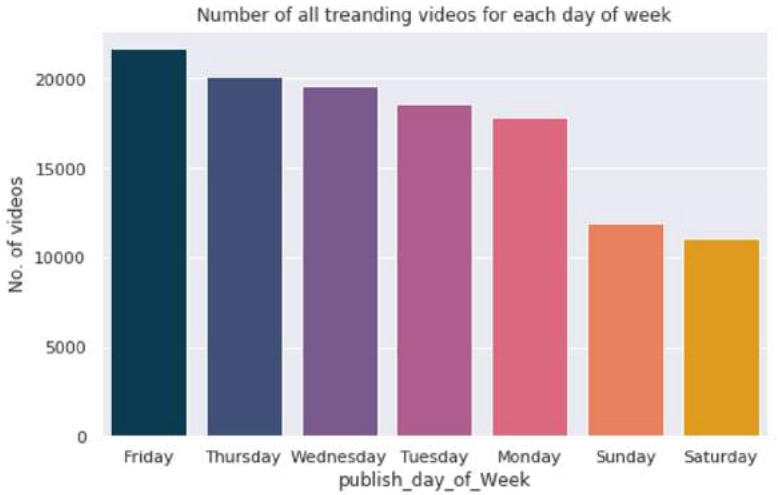
(120746, 18)

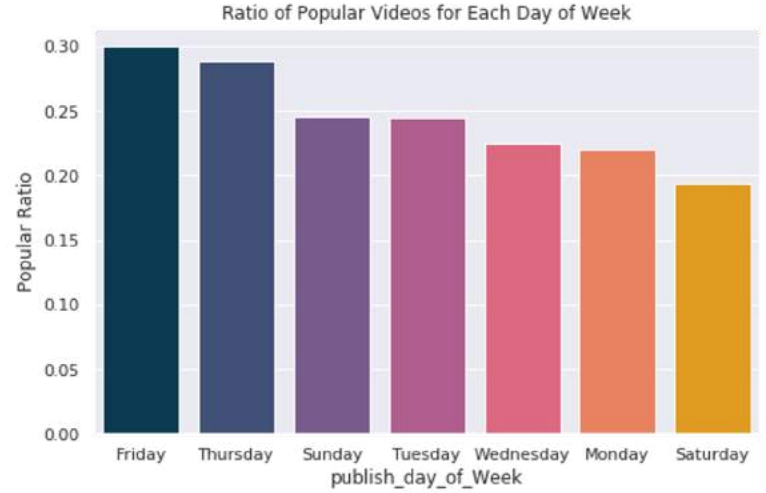
Target Variable: views_cat

(120746, 19)

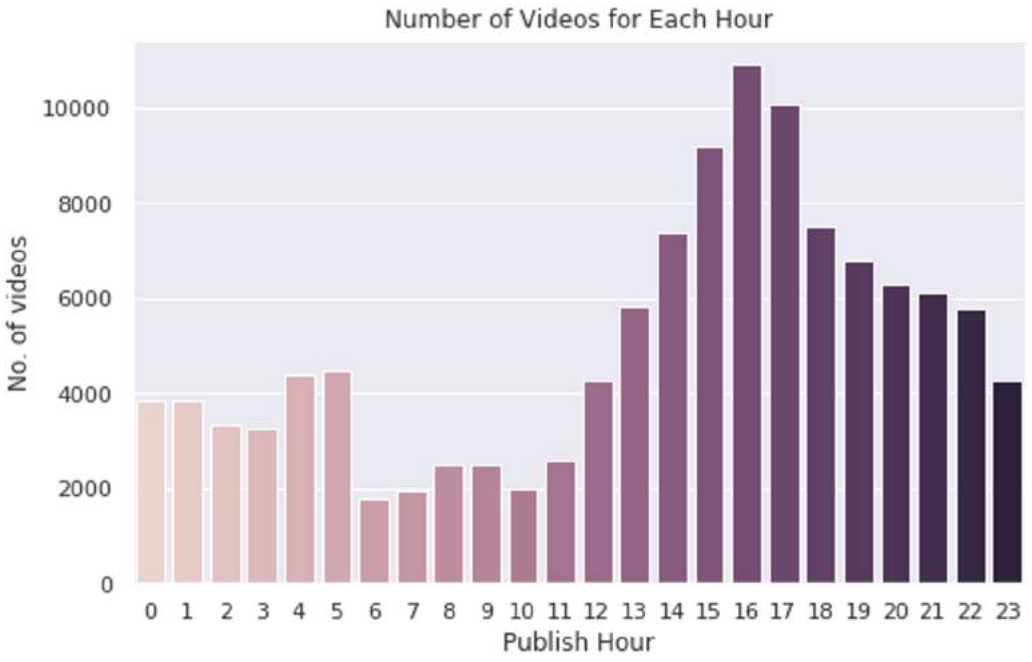
Publish Time: publish_day_of_Week

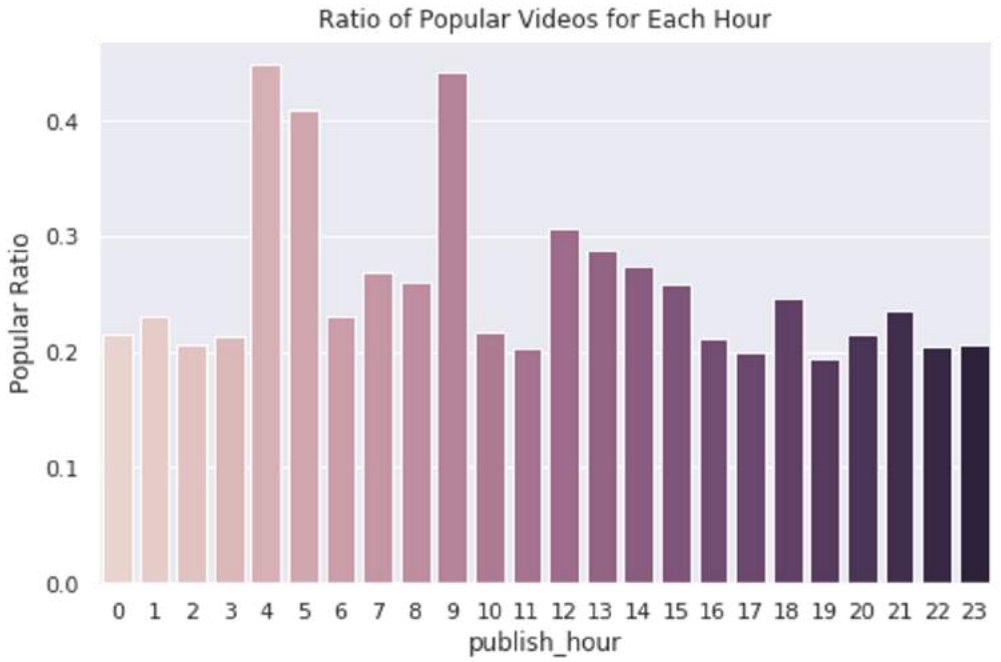
- Create new feature called “publish_day_of_week” from “publish_time”.
- Separate date, hour and time into three columns from 'publish_time' column.
- Create feature to calculate the number of days between publish date and trending date.



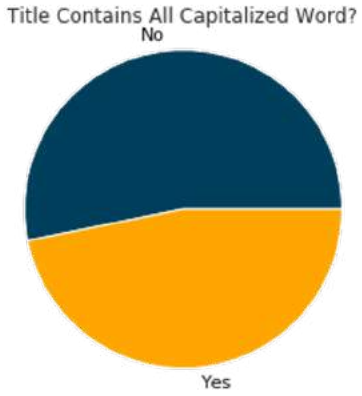


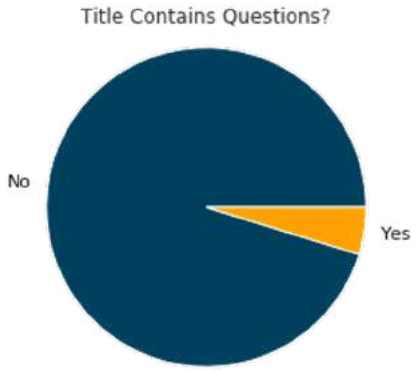
Publish Time: publish_hour





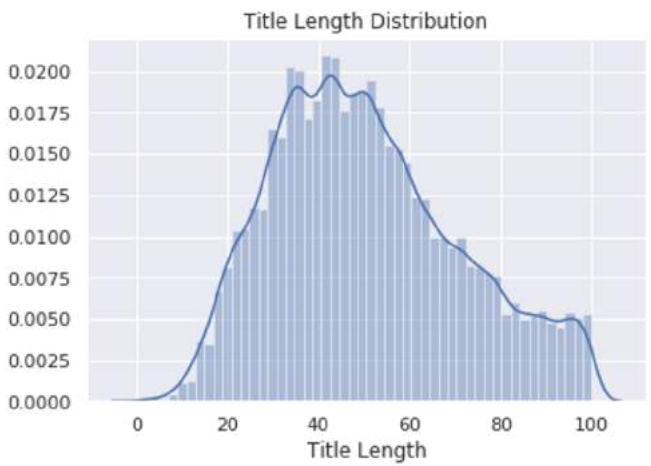
Titles and Tags

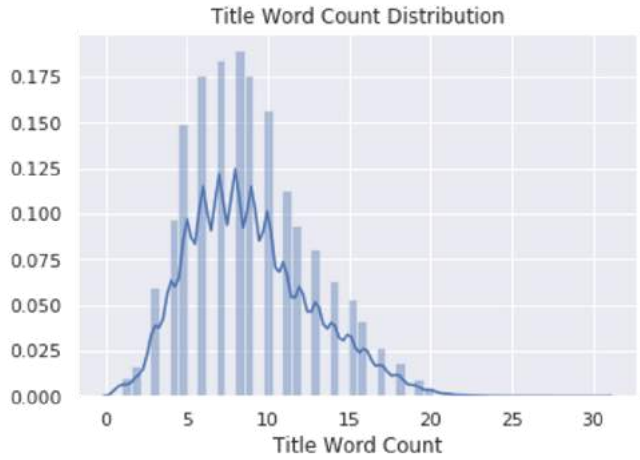


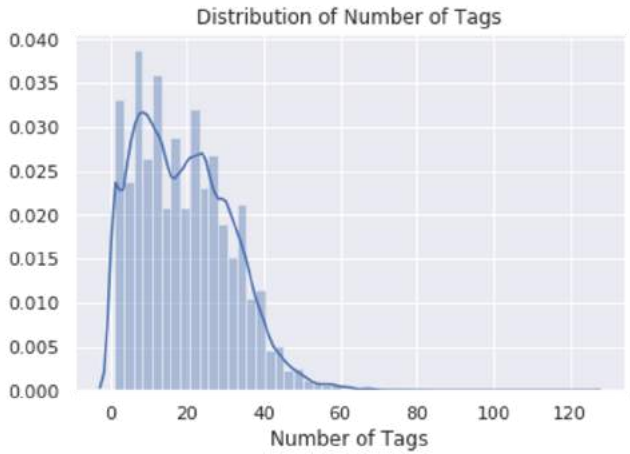


We create new features from titles and tags:

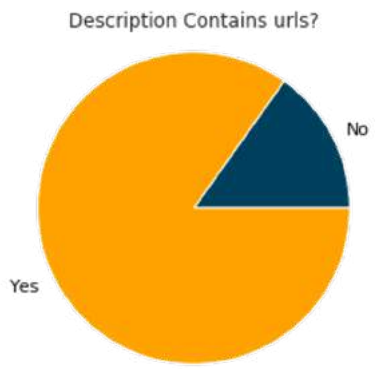
- title_char_length
- title_word_count
- title_allcap
- title_question
- tag_nums

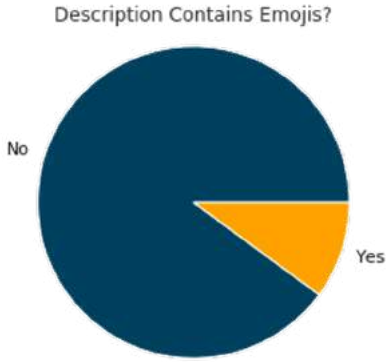


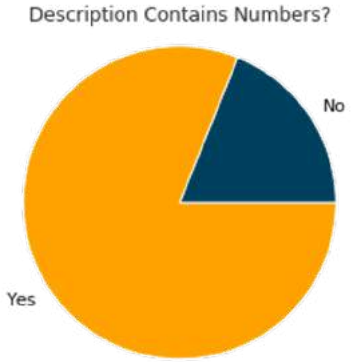


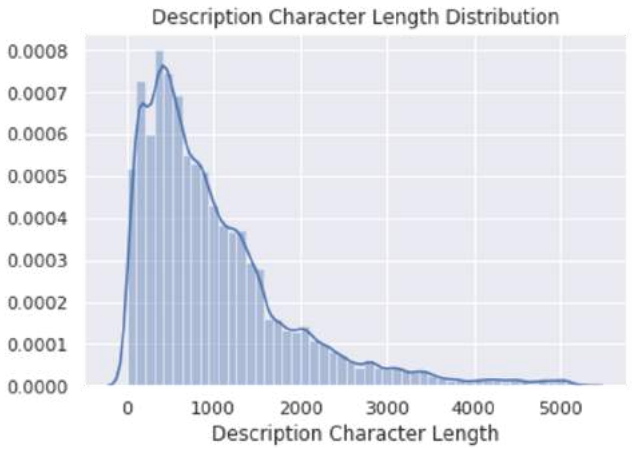


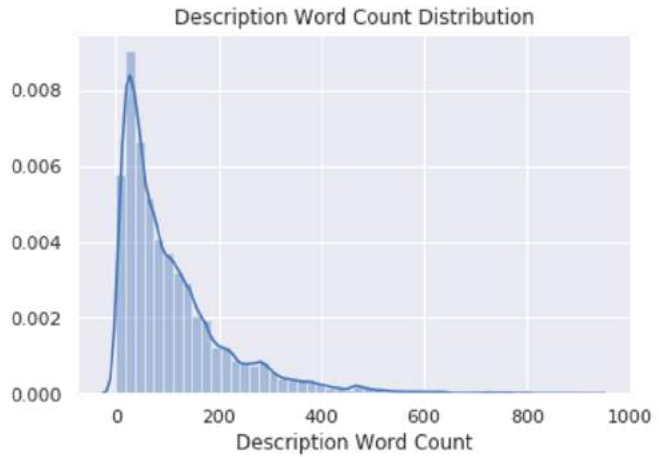
Descriptions









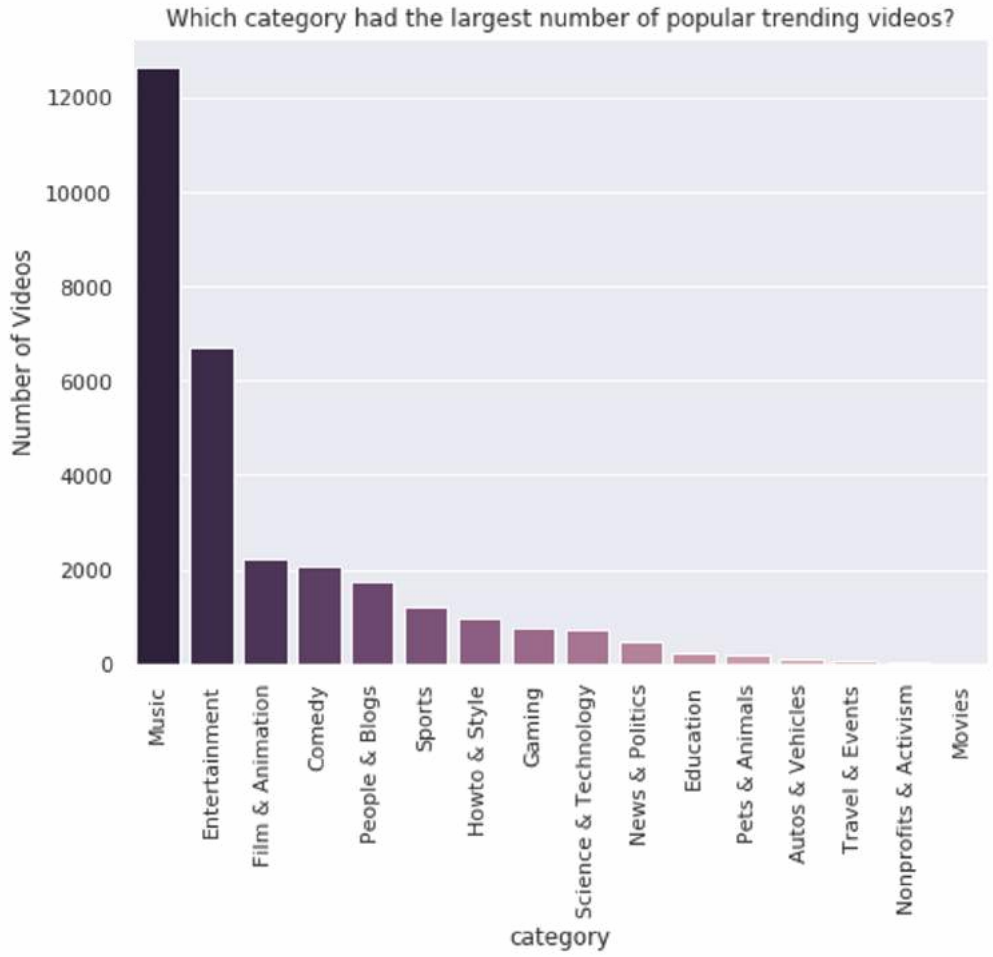


We create new features from descriptions::

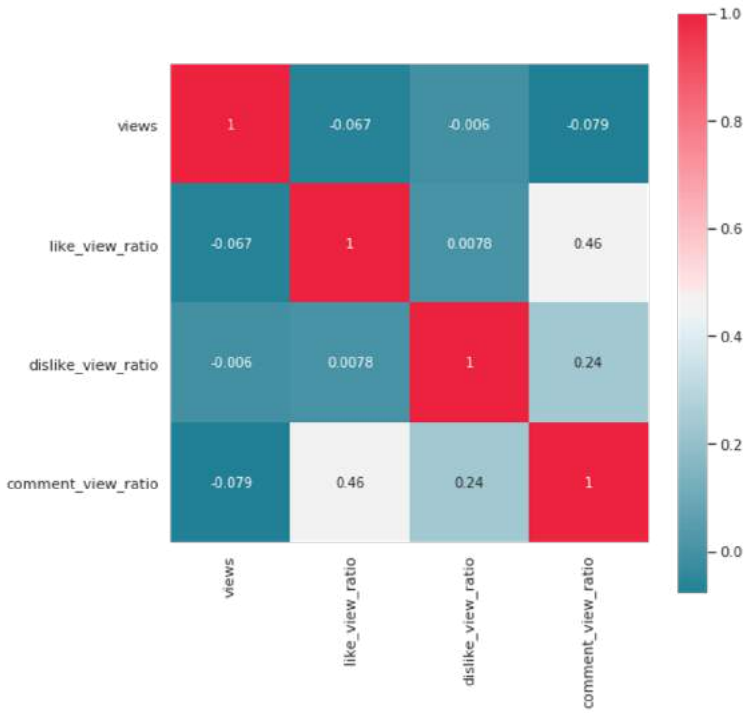
desc_urls
desc_emojis
desc_nums
desc_char_length
desc_word_count

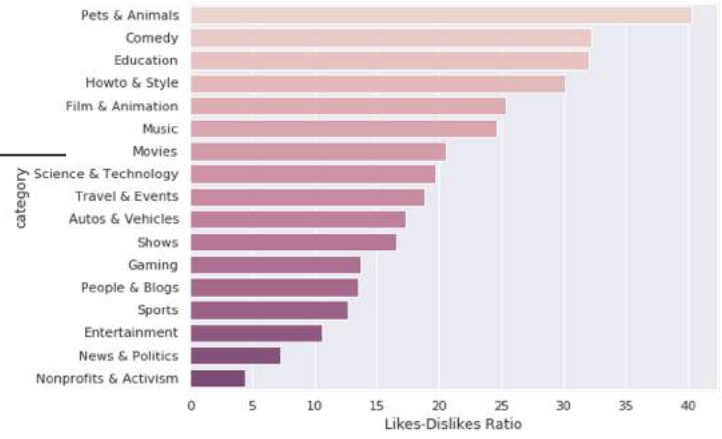
Popular Videos and Categories

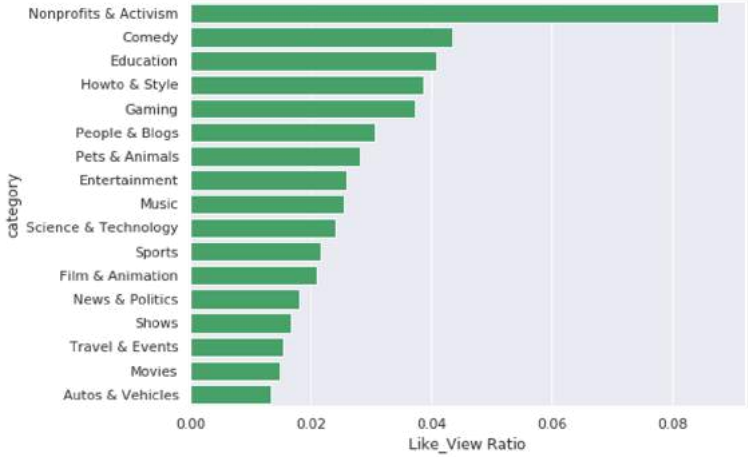
Music	12636
Entertainment	6686
Film & Animation	2230
Comedy	2054
People & Blogs	1751
Sports	1212
Howto & Style	974
Gaming	776
Science & Technology	728
News & Politics	480
Education	225
Pets & Animals	206
Autos & Vehicles	106
Travel & Events	59
Nonprofits & Activism	9
Movies	4



Views, likes, dislikes and comments ratio





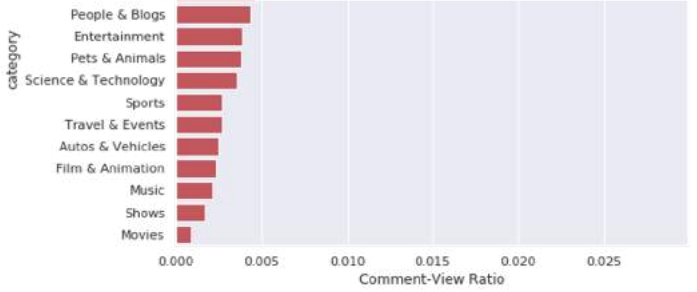


Suicide: Be Here Tomorrow. - YouTube

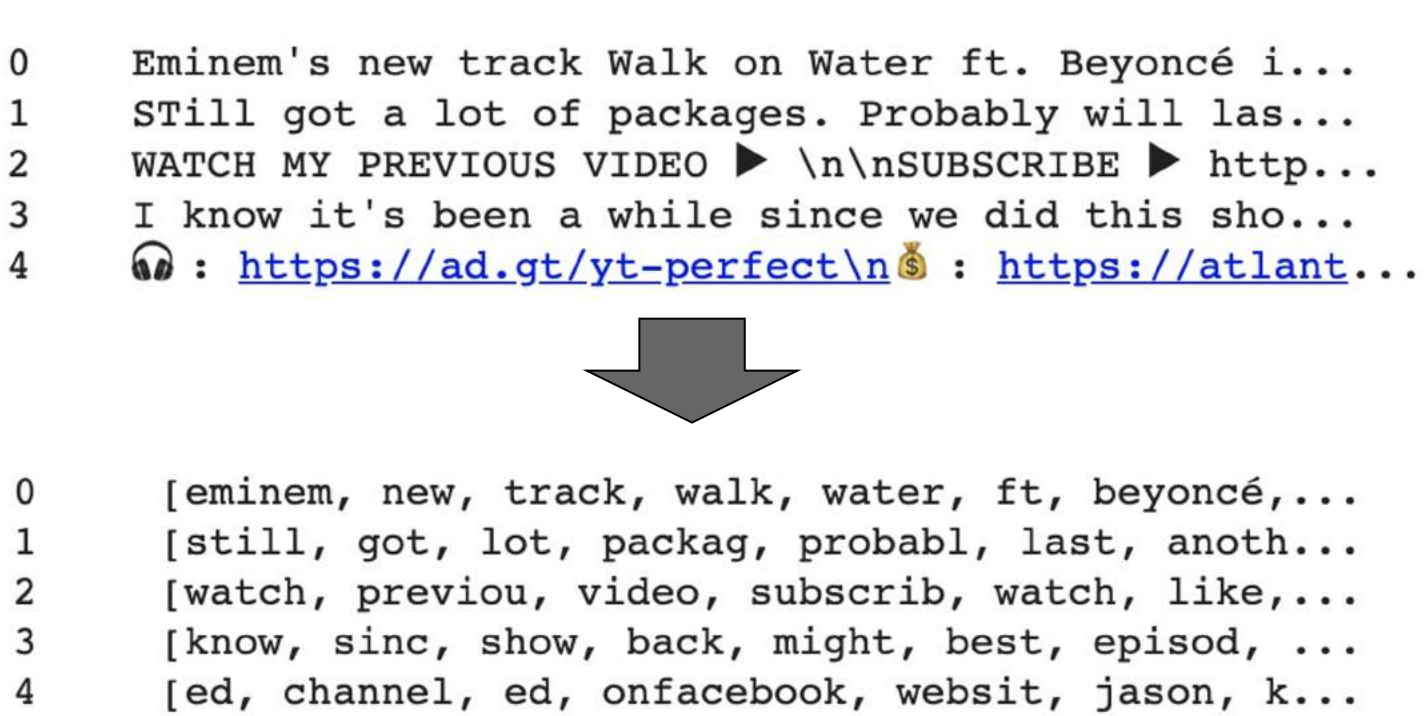
YouTube · Logan Paul

Яхты, олигархи, девочки: охотница на мужчин разоблачает ...

YouTube · Алексей Навальный



Text Data Cleaning



Covert Description

Convert to lowercase;

Remove numbers, punctuations, emojis, urls, leading/ending spaces

Basic NLP

- Tokenization
- Stemming
- Lemmatization

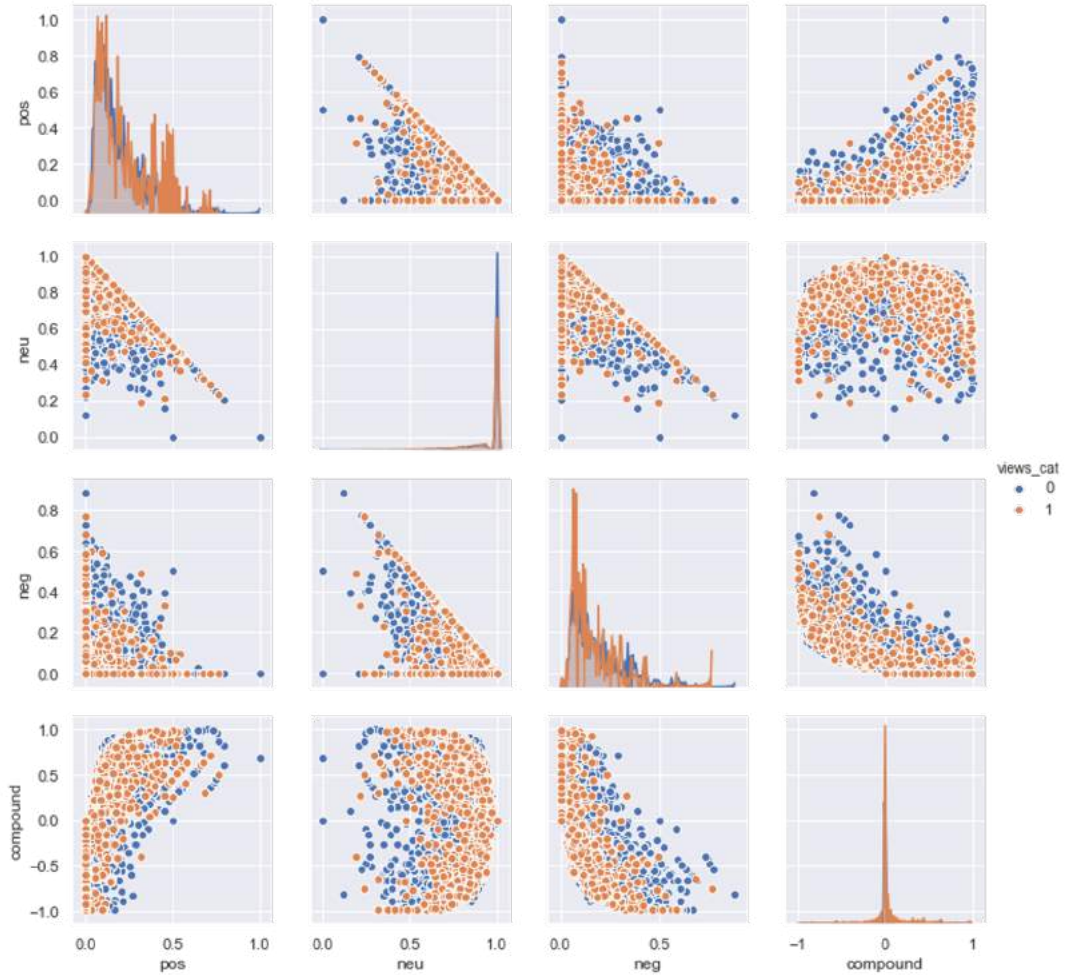
Sentiment Analysis for Tags

	country	tags	video_id	neg	neu	pos	compound
0	CA	Eminem|"Walk"|"On"|"Water"|"Aftermath/Shady/In...	n1WpP7iowLc	0.0	1.0	0.0	0.0000
1	CA	plush|"bad unboxing"|"unboxing"|"fan mail"|"id...	0dBIkQ4Mz1M	0.0	1.0	0.0	0.0000
2	CA	racist superman|"rudy"|"mancuso"|"king"|"bach"...	5qpjK5DgCt4	0.2	0.8	0.0	-0.6124
3	CA	ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...	d380meD0W0M	0.0	1.0	0.0	0.0000
4	CA	edsheeran|"ed sheeran"|"acoustic"|"live"|"cove...	2Vv-BfVoq4g	0.0	1.0	0.0	0.0000

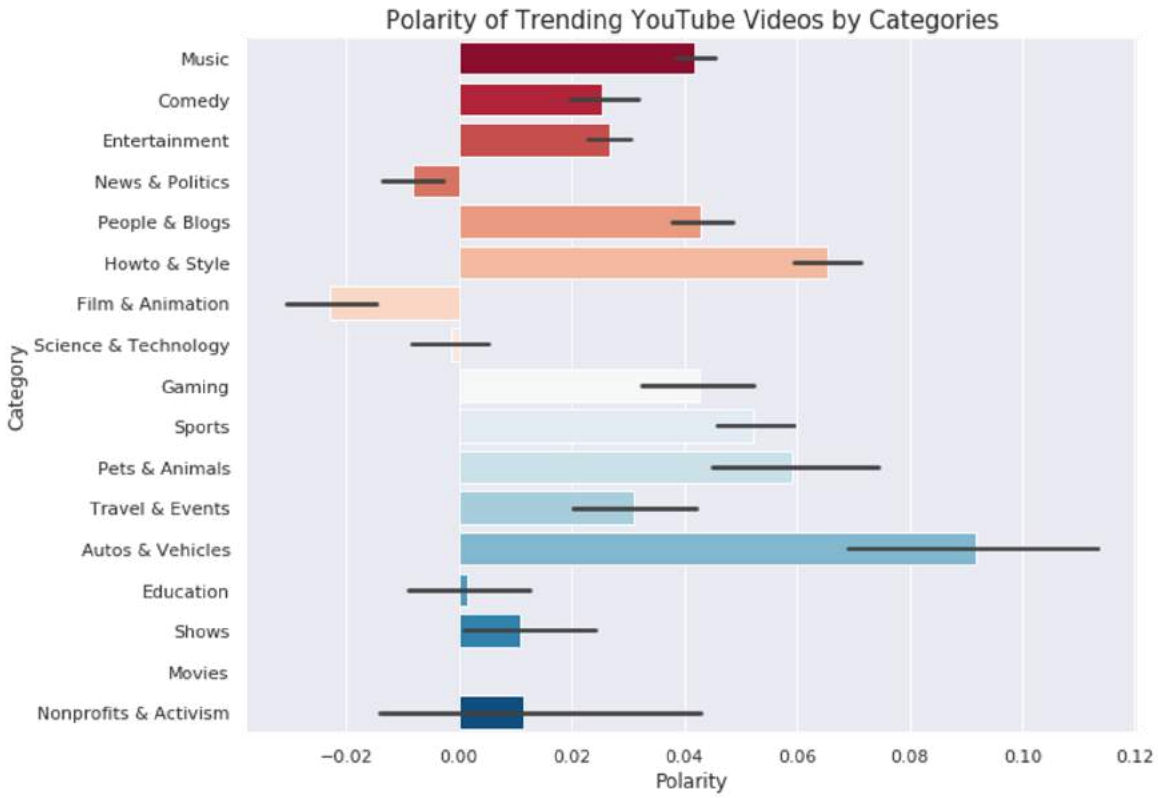
Compound Score - sum of all the lexicon ratings which have been normalized between -1 and +1

positive sentiment: c.s. >= 0.05
neutral sentiment: -0.05 < c.s. < 0.05
negative sentiment: c.s. <= -0.05

Pairplot of Scores by views_cat



Polarity of Videos by Category

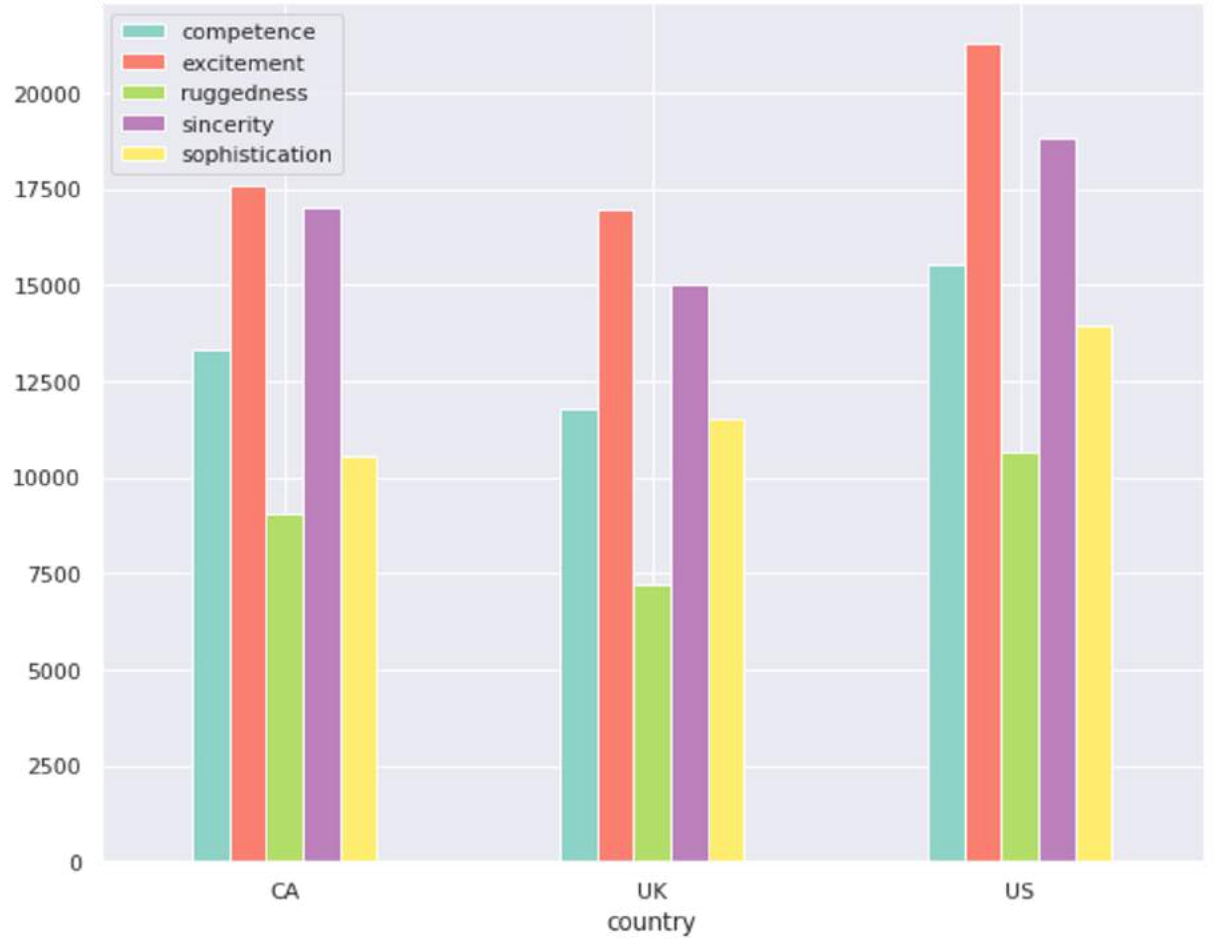


Film & Animation

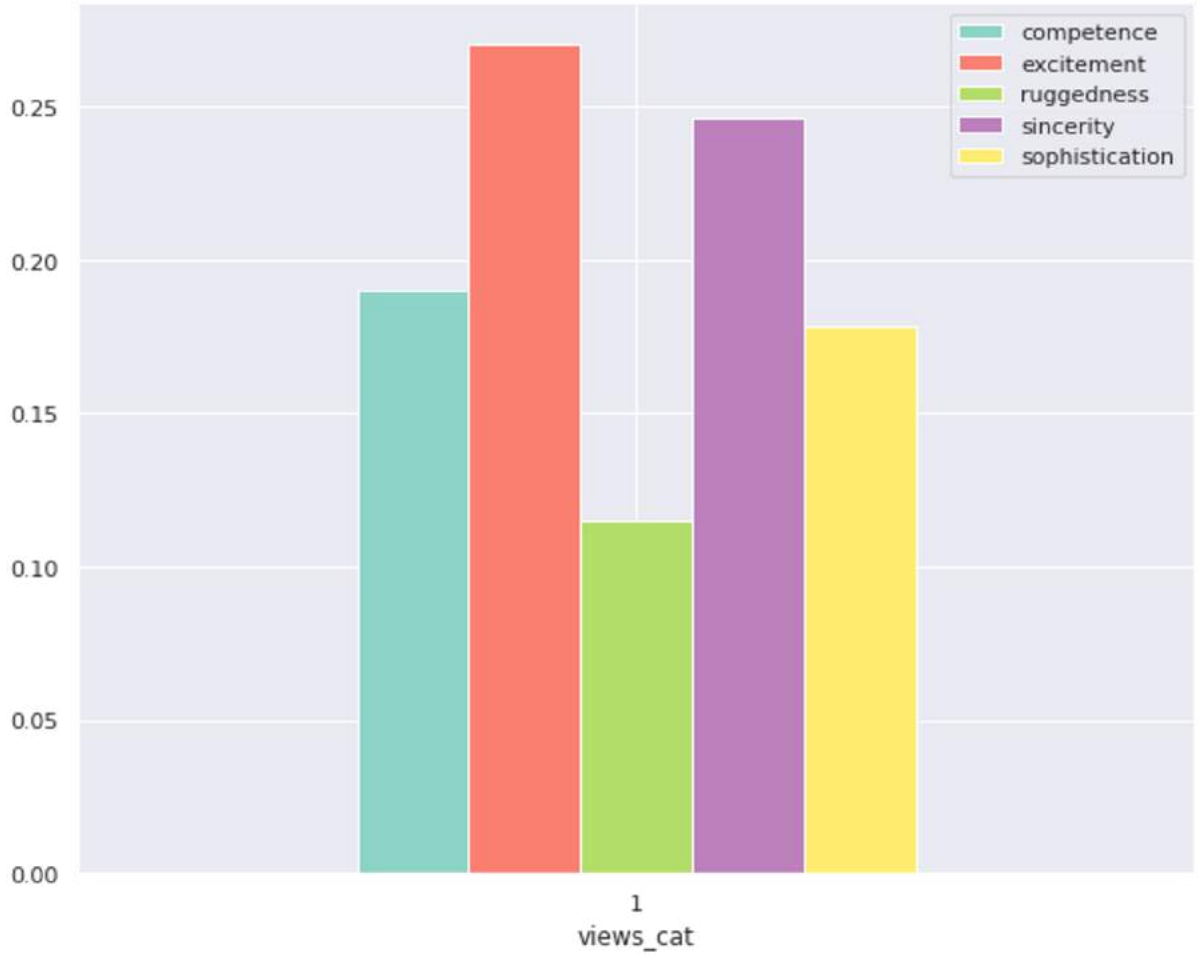
Autos & Vehicles

Identify Video's Brand Personality

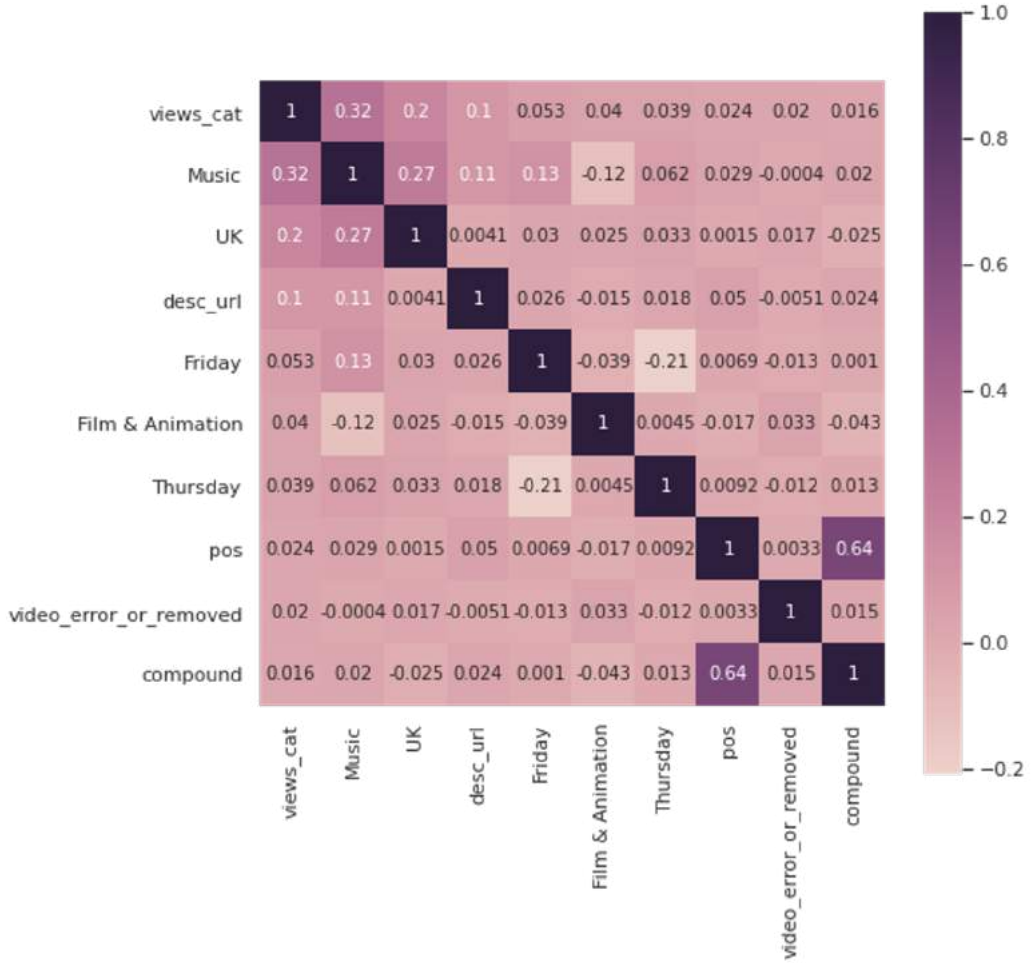
	Personality	Word
0	COMPETENCE	ABLE (1)
1	COMPETENCE	ABLE_BODIED (1)
2	COMPETENCE	ADEPT (1)
3	COMPETENCE	ADROIT (1)
4	COMPETENCE	ASSIDUOUS (1)
5	COMPETENCE	ASSURED (1)
6	COMPETENCE	ASTUTE (1)
7	COMPETENCE	AWARD_WINNING (1)
8	COMPETENCE	BLOOMING (1)
9	COMPETENCE	BOOMING (1)



Video’s Personality by views_cat



Correlation Matrix with Top 10 Features



Initial Dataset

shape: (120746, 17)

Current Dataset

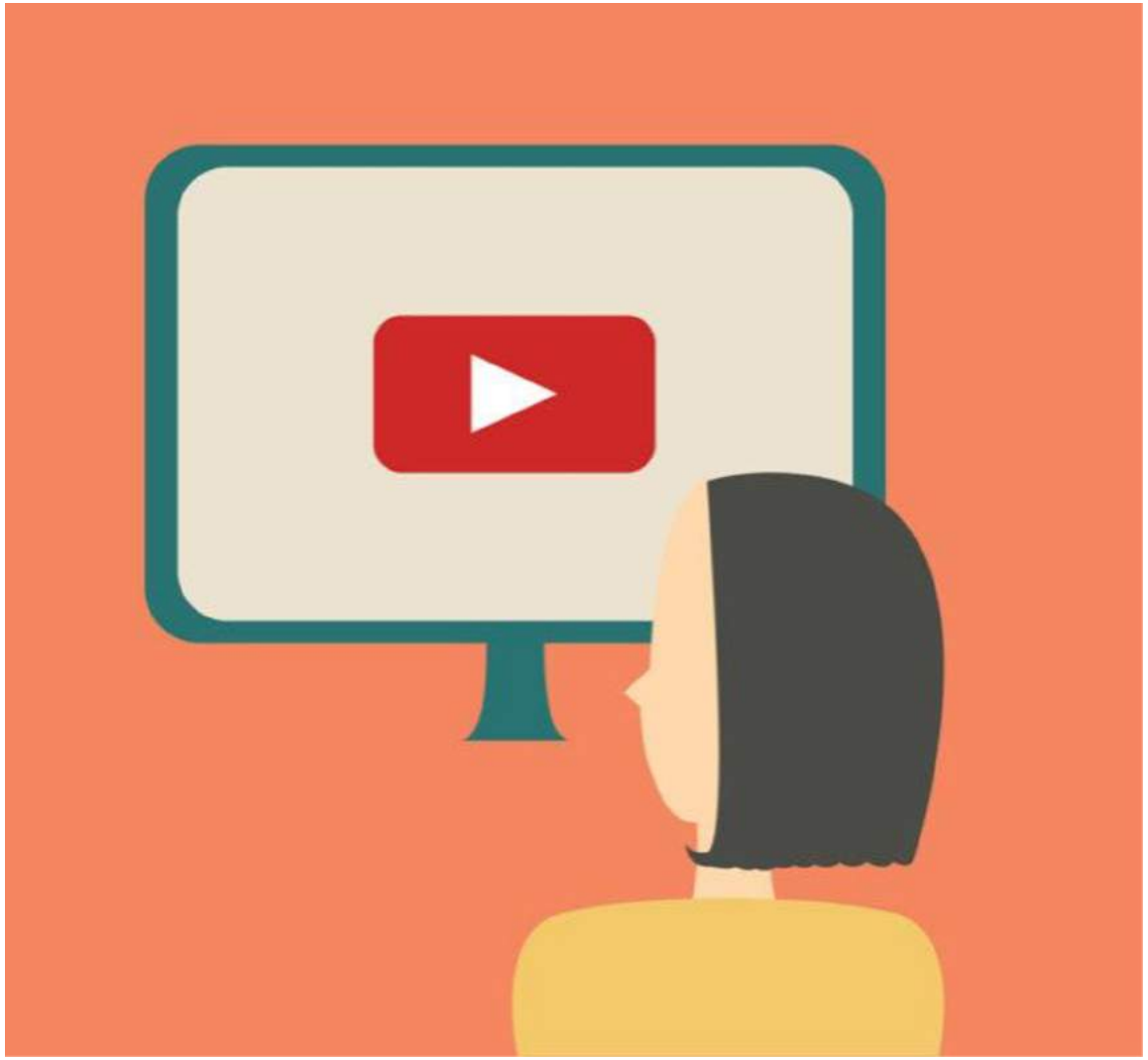
(after Feature Engineering)

shape: (120746, 78)

Final Dataset

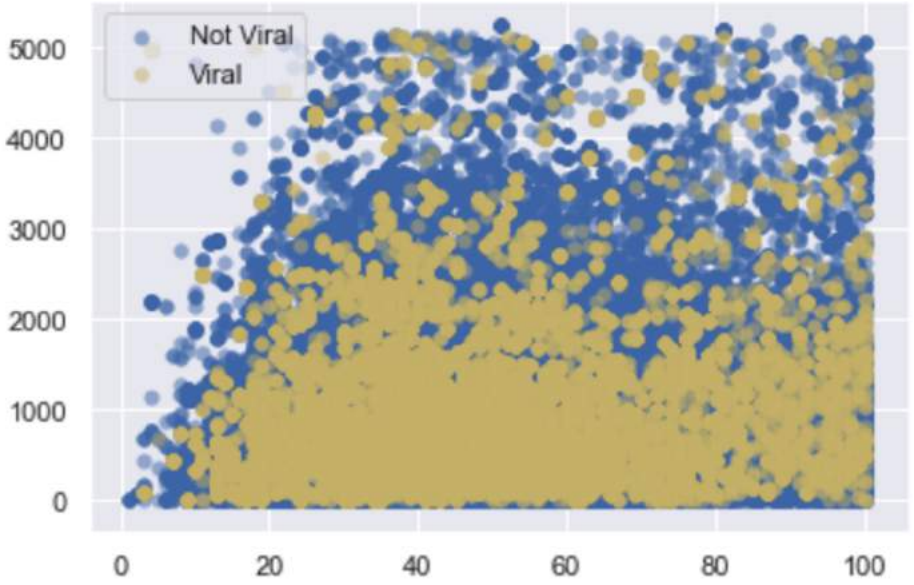
(after dropping object/repeated features)

shape: (120746, 51)

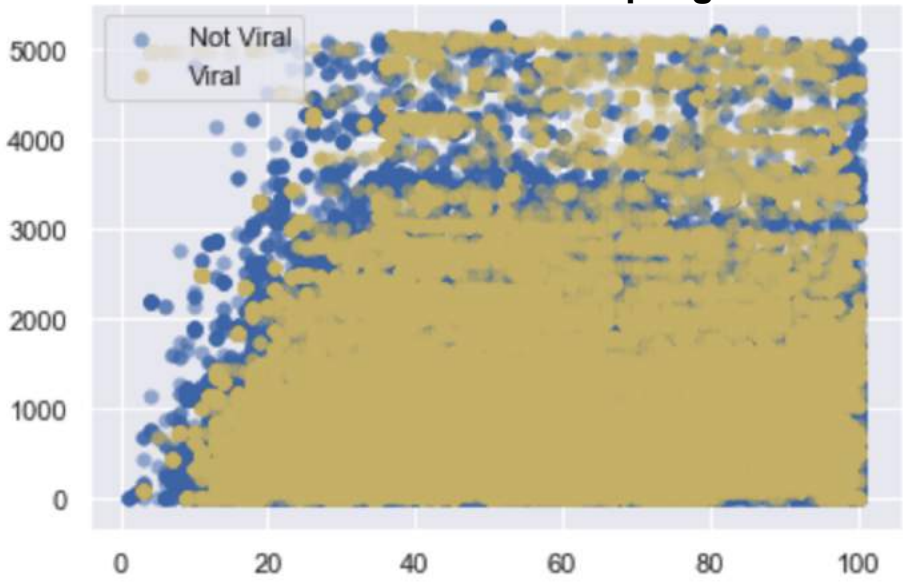


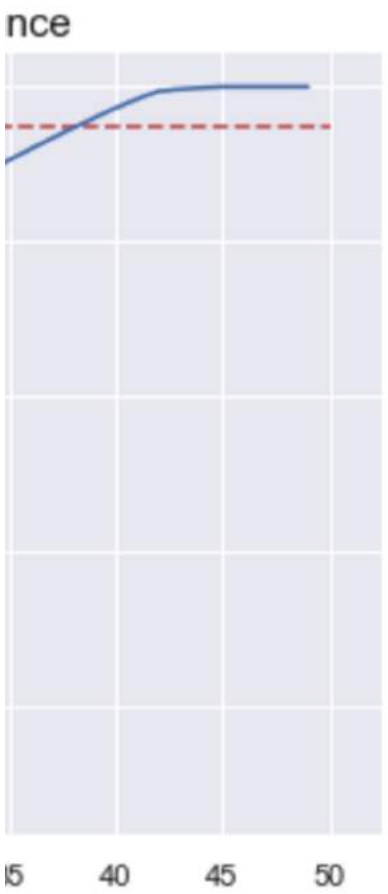
Modeling and Prediction

Original Dataset



Dataset after Oversampling





Standardization

imension Reduction: PCA

versampling using Adaptive Synthetic Sampling (ADASYN)

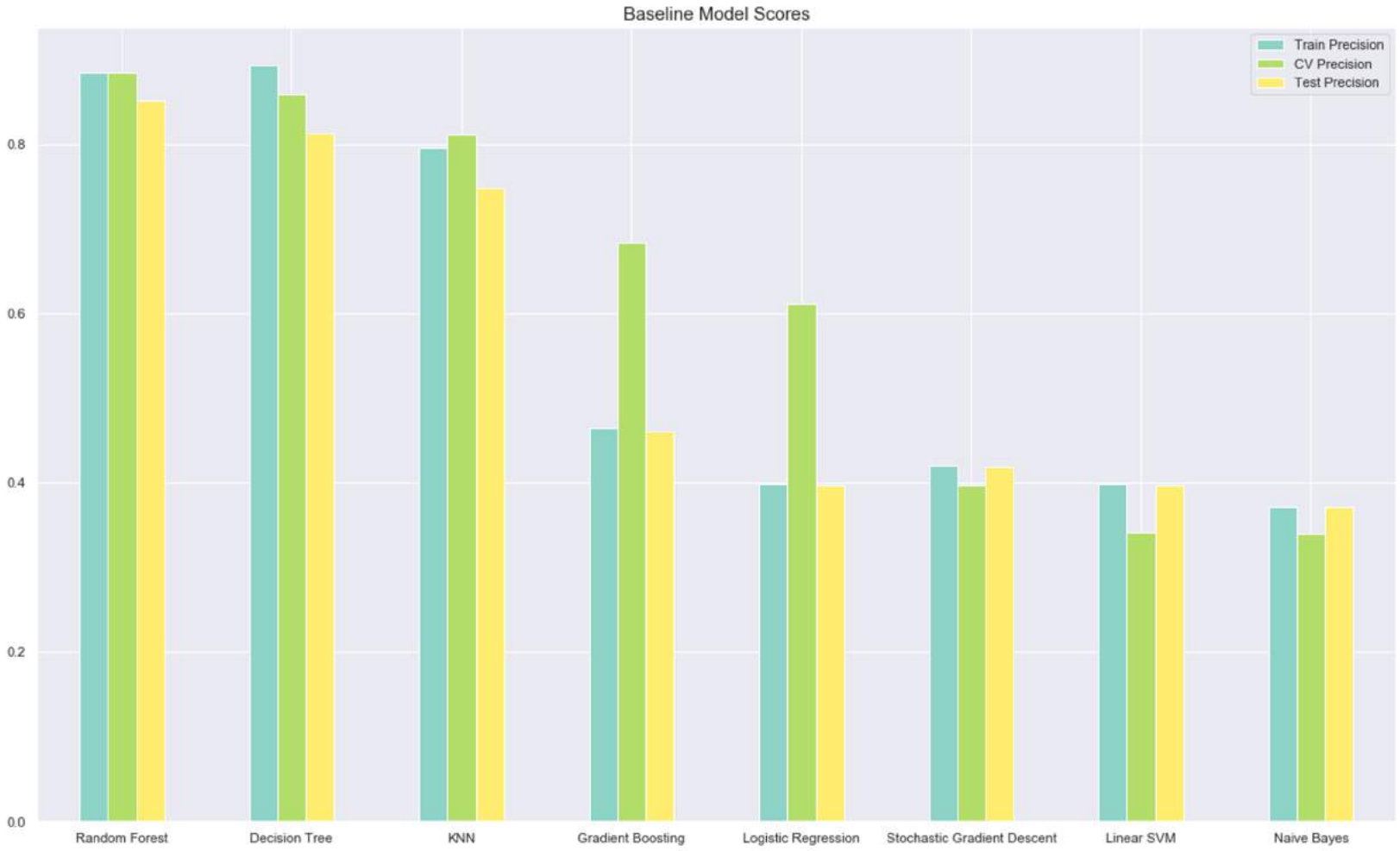
Modeling

ort fier	Stochastic Gradient Descent Classifier	K Nearest Neighbor	Decision Tree Classifier	Random Forest Classifier	Gradient Boosting Classifier

Hyperparameter Tuning: Random Search

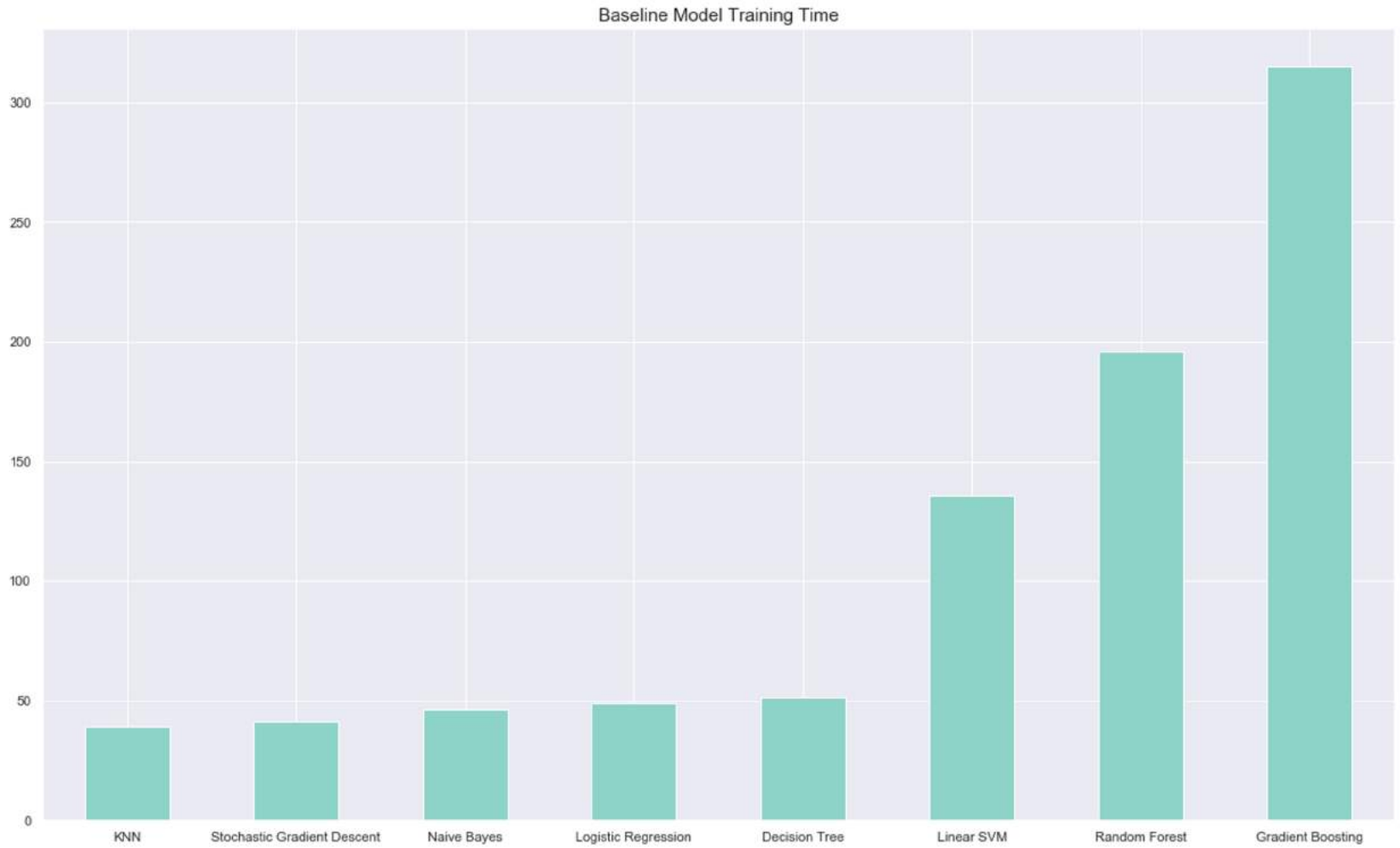
Precision

Baseline Model



Training Time

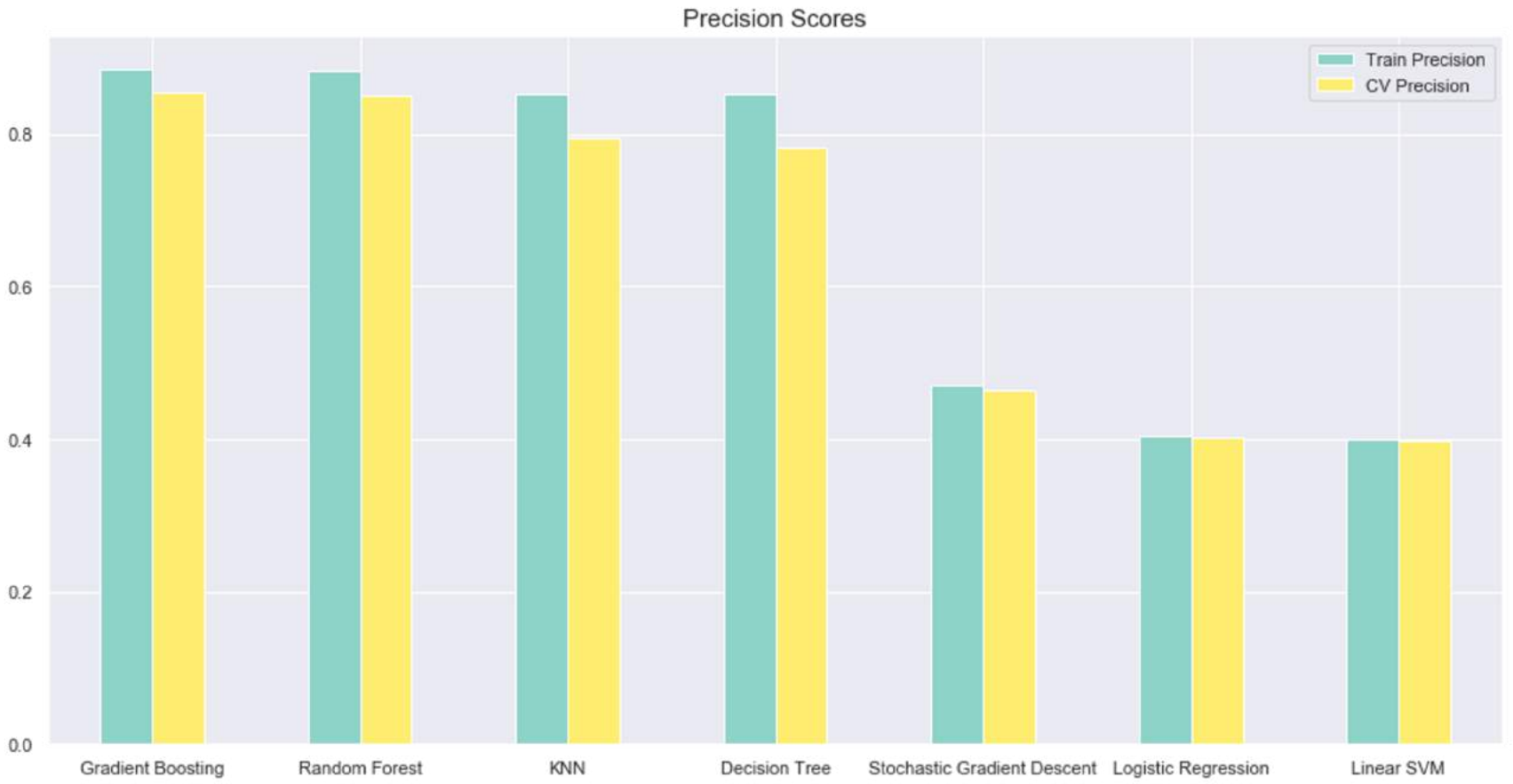
Baseline Model



Precision

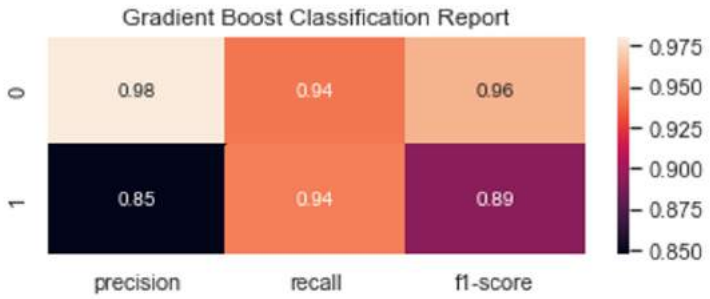
Hyperparameter Tuning

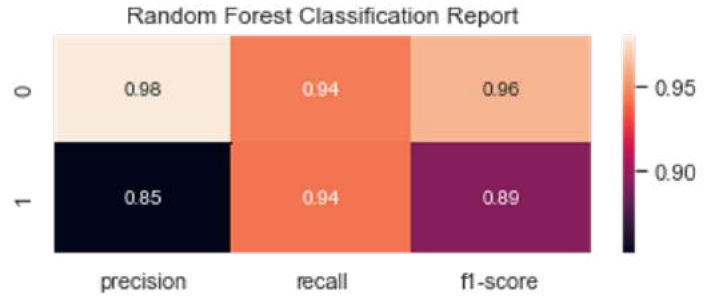
- Choose Gradient Boosting Classifier as the final model
- Compare training score and cross validation score to check overfitting/underfitting

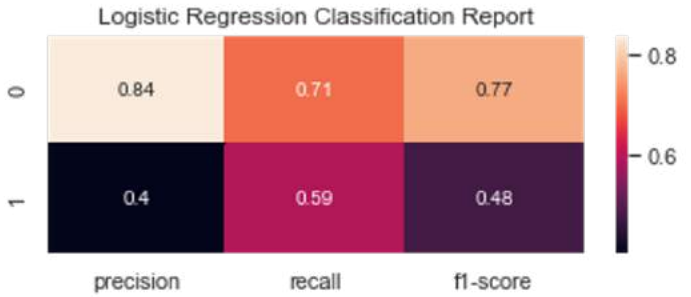


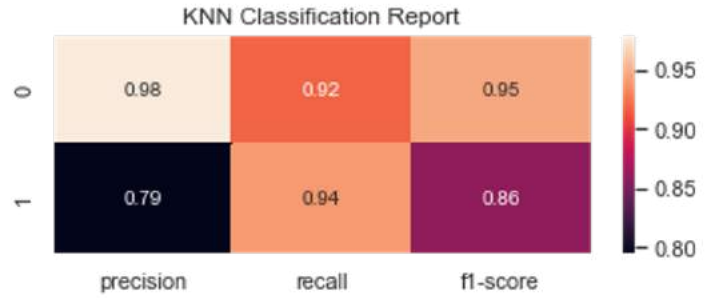
Classification Report

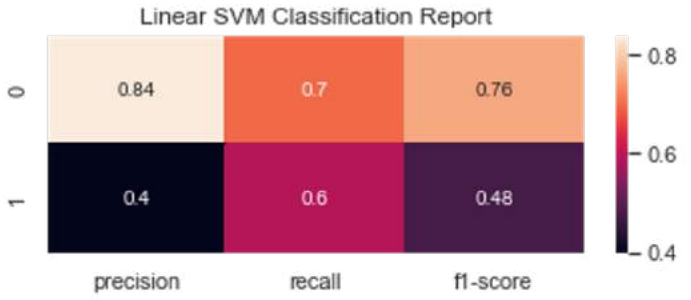
Hyperparameter Tuning

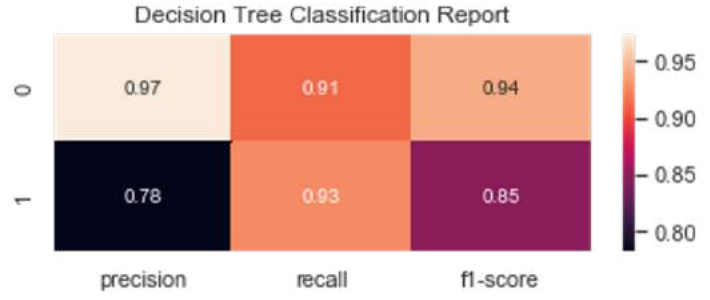


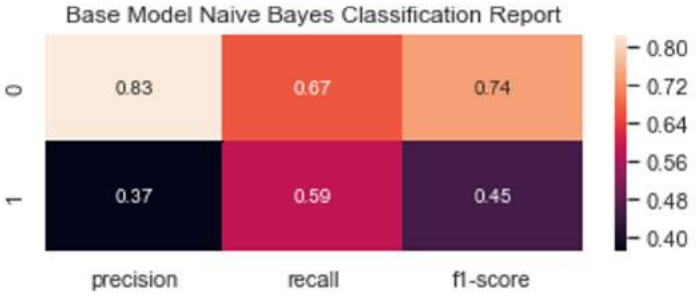








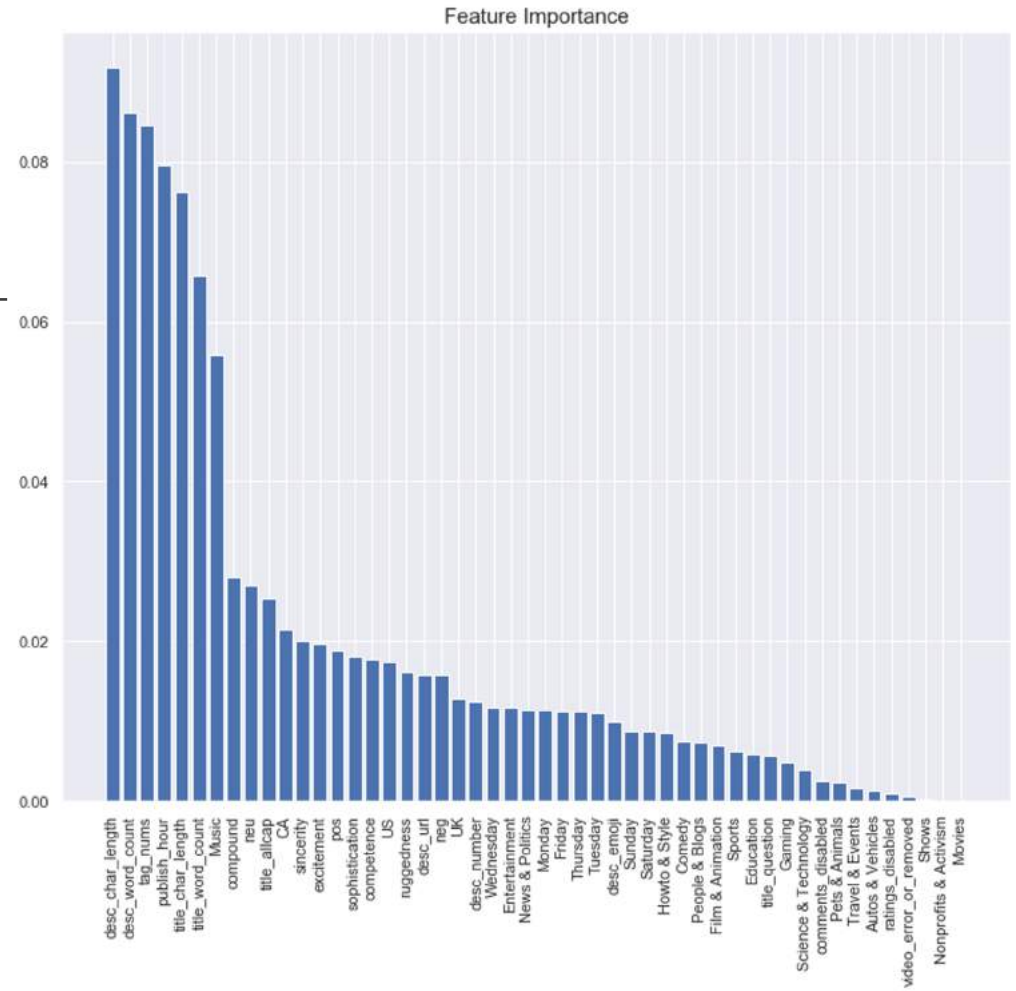




Feature Importance

Most important features

- Description length
- Number of tags
- Title length
- Publish hour
- Music



comments_disabled
ratings_disabled
video_error_or_removed
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
publish_hour
title_char_length
title_word_count
title_allcap
title_question
tag_nums
desc_url
desc_emoji
desc_number
desc_char_length
desc_word_count
competence
excitement
ruggedness
sincerity
sophistication
neg

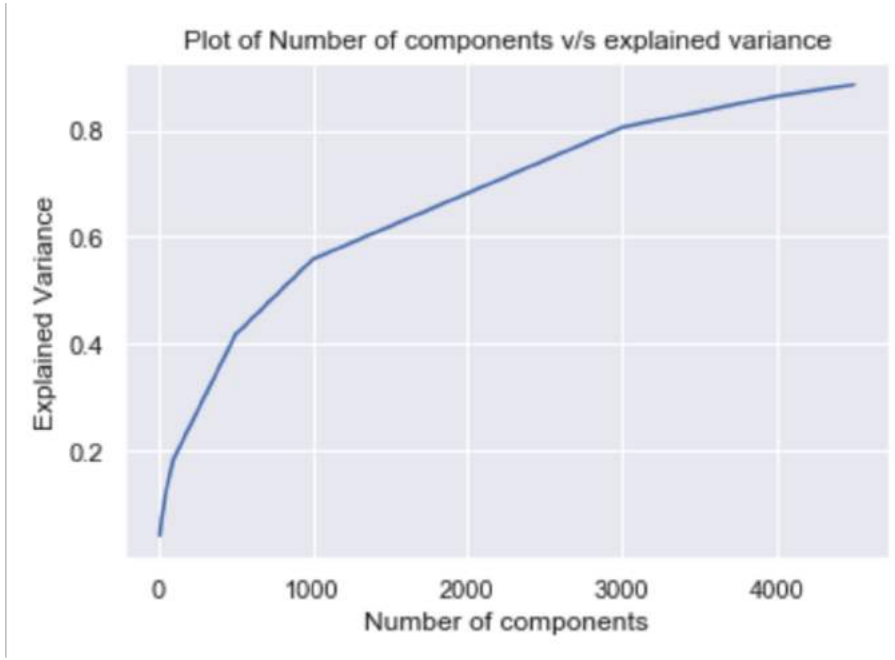
neu
pos
compound
Autos & Vehicles
Comedy
Education
Entertainment
Film & Animation
Gaming
Howto & Style
Movies
Music
News & Politics
Nonprofits & Activism
People & Blogs
Pets & Animals
Science & Technology
Shows
Sports
Travel & Events
CA
UK
US

YouTube

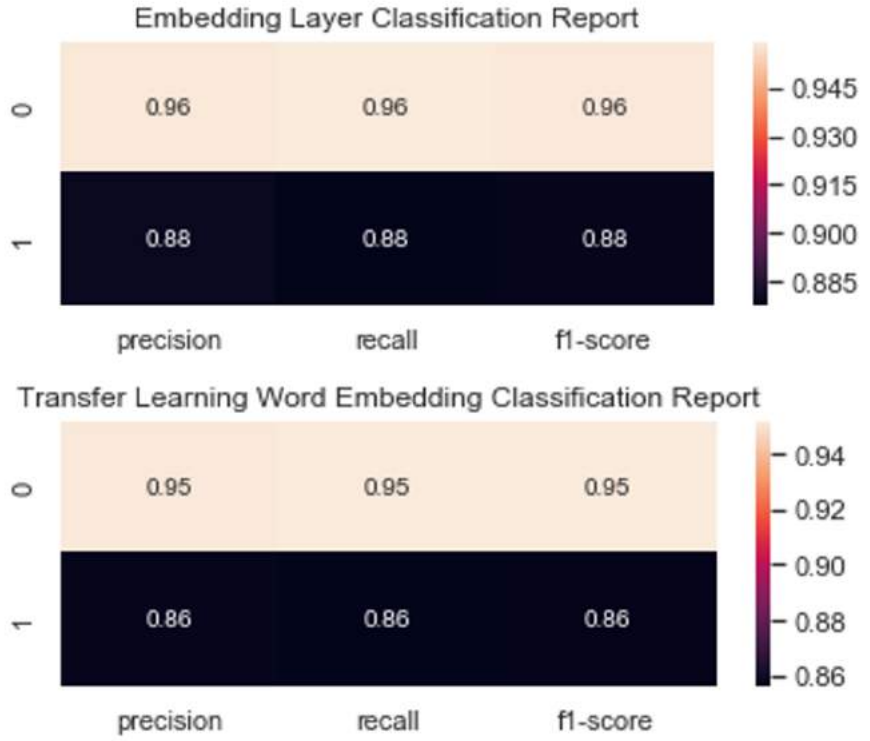
Unstructured Data Modeling

Tfidf Vectorizer
Dimension: 46,433

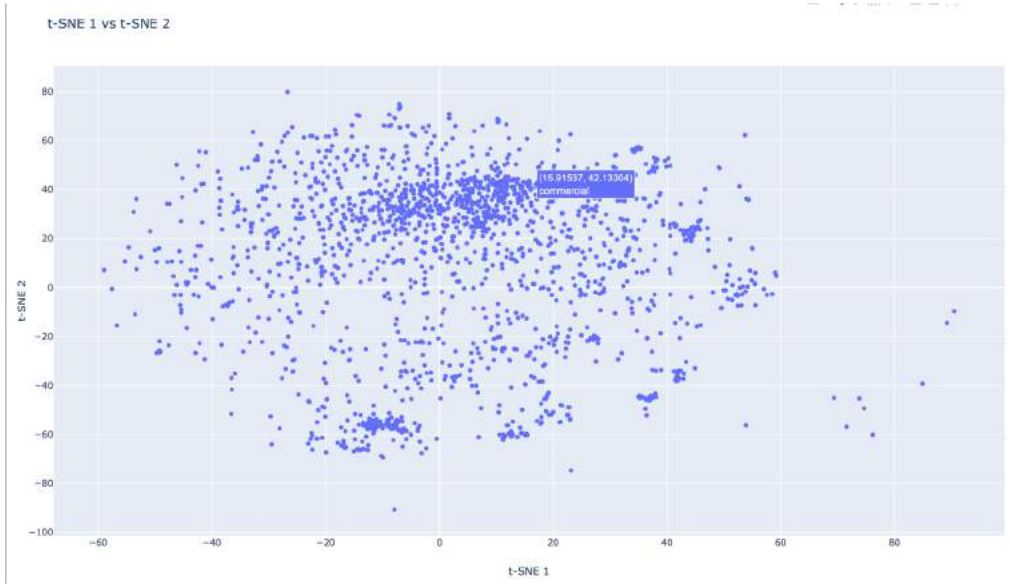
TruncatedSVD
Dimension: 4,000 (85% Explained Variance)



Word Embedding

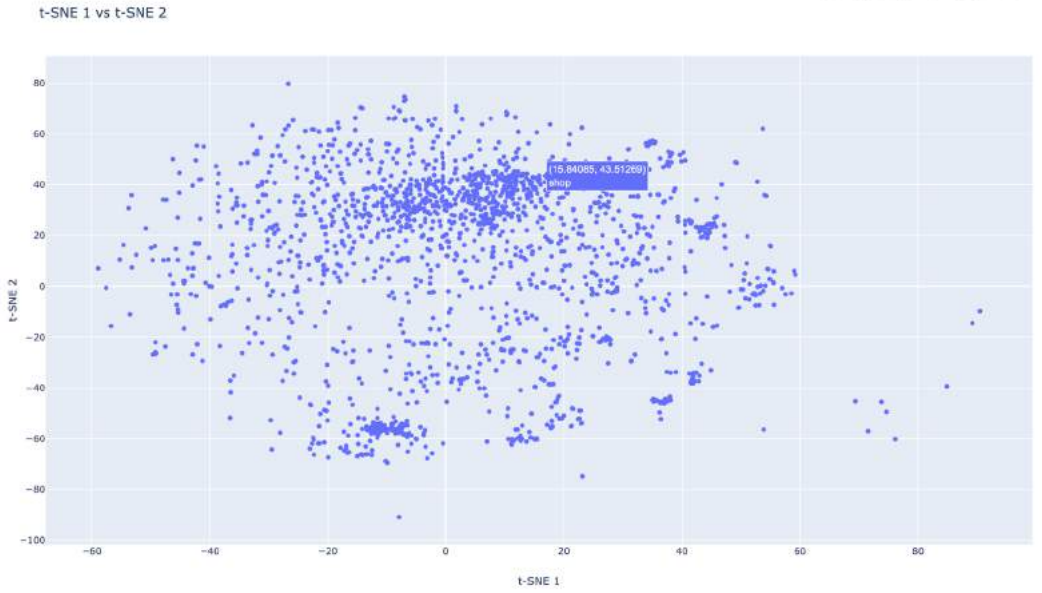


Word Vector Space



Commercial

Shop



Future Work and Deployment

Project Potential

- Current Project

Classification model for marketing companies to spot videos most likely to go viral

- Potential Improvement

A recommendation system that matches video suggestion with company's brand personality

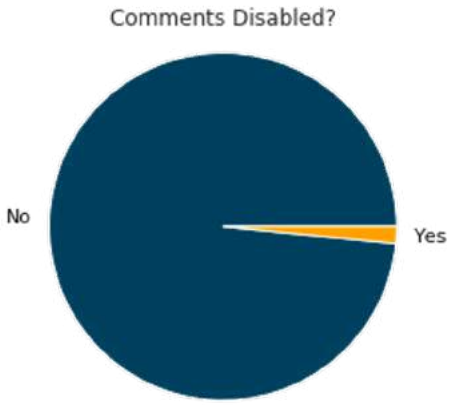
- Future Development

Analyze company's target clientele and recommend videos with matching target audience

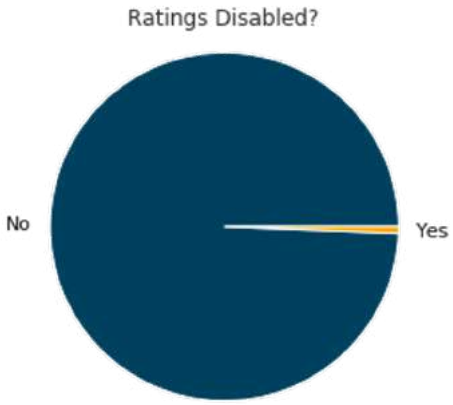
Thank you!

Appendix: Three Boolean Features

False 120627
True 119
Name:
video_error_or_removed,
dtype: int64

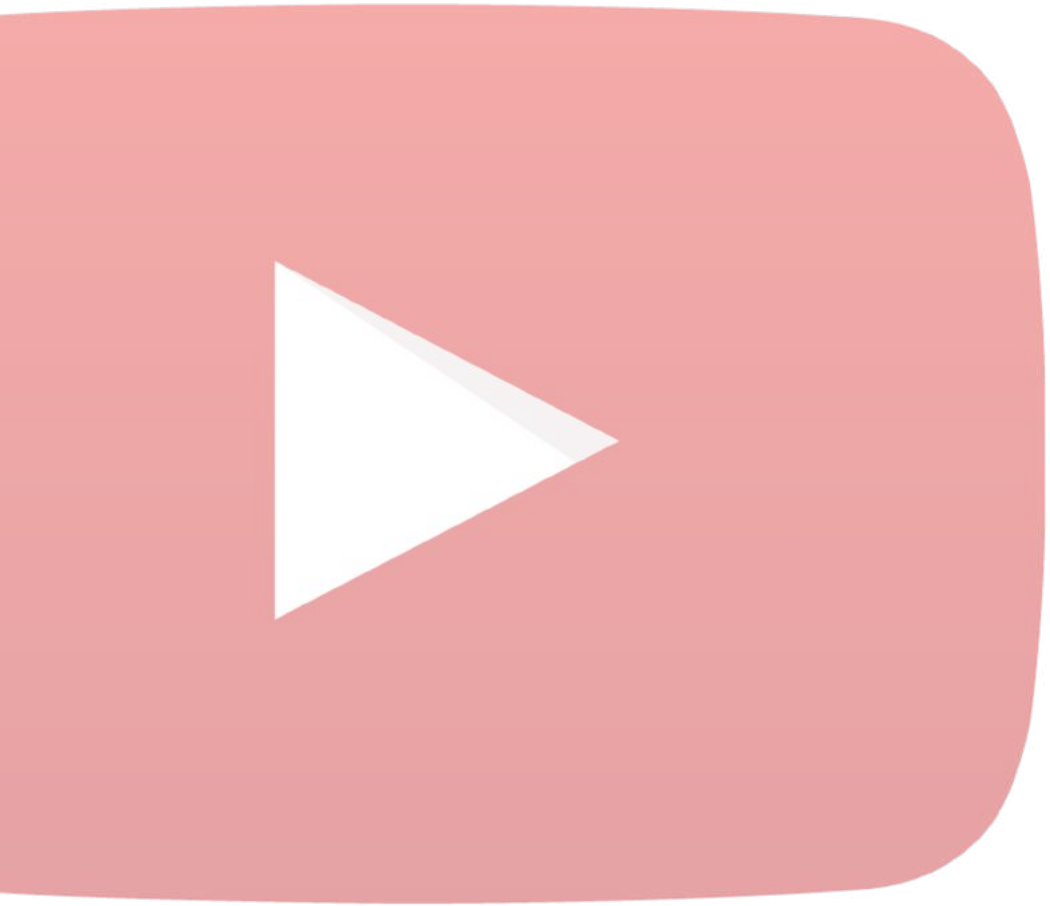


False 118847
True 1899
Name:
comments_disabled,
dtype: int64



False 120026
True 720
Name: ratings_disabled,
dtype: int64

Appendix: Final Model



Gradient Boosting Classifier

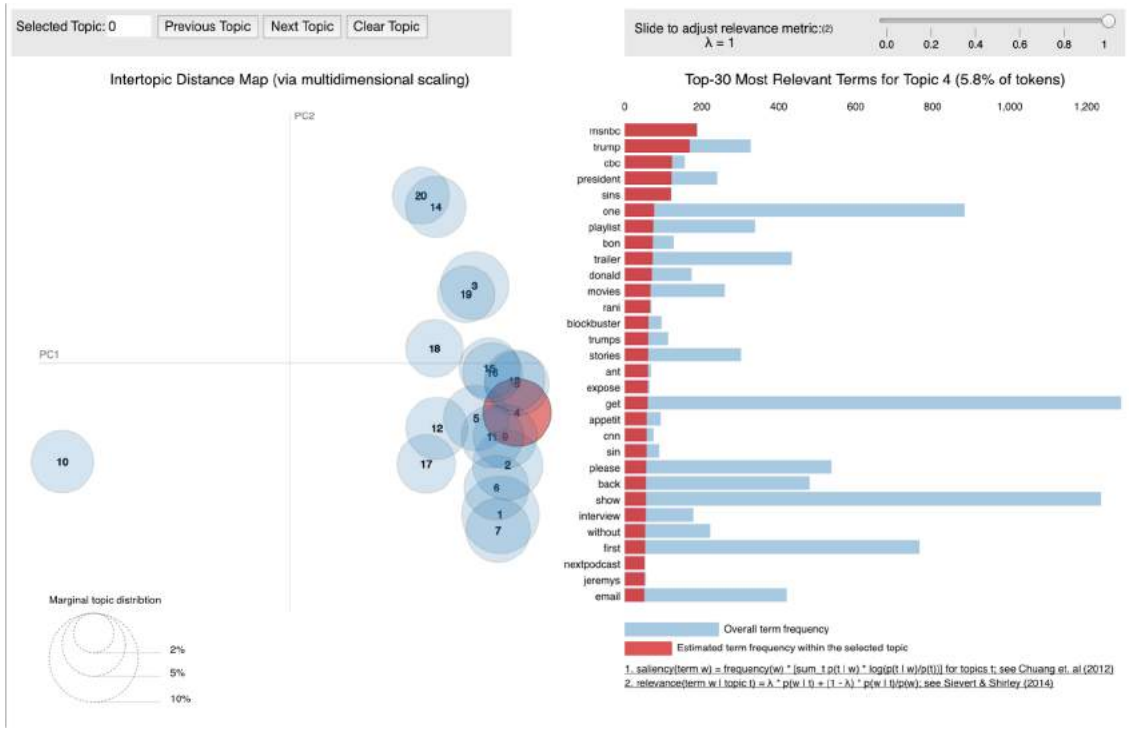
- 'learning_rate': 0.001
- 'max_depth': 59
- 'max_features': 7
- 'min_samples_leaf': 4
- 'min_samples_split': 8
- 'n_estimators': 186
- 'subsample': 1.0

Appendix:
Steps for Building the Embedding Layer

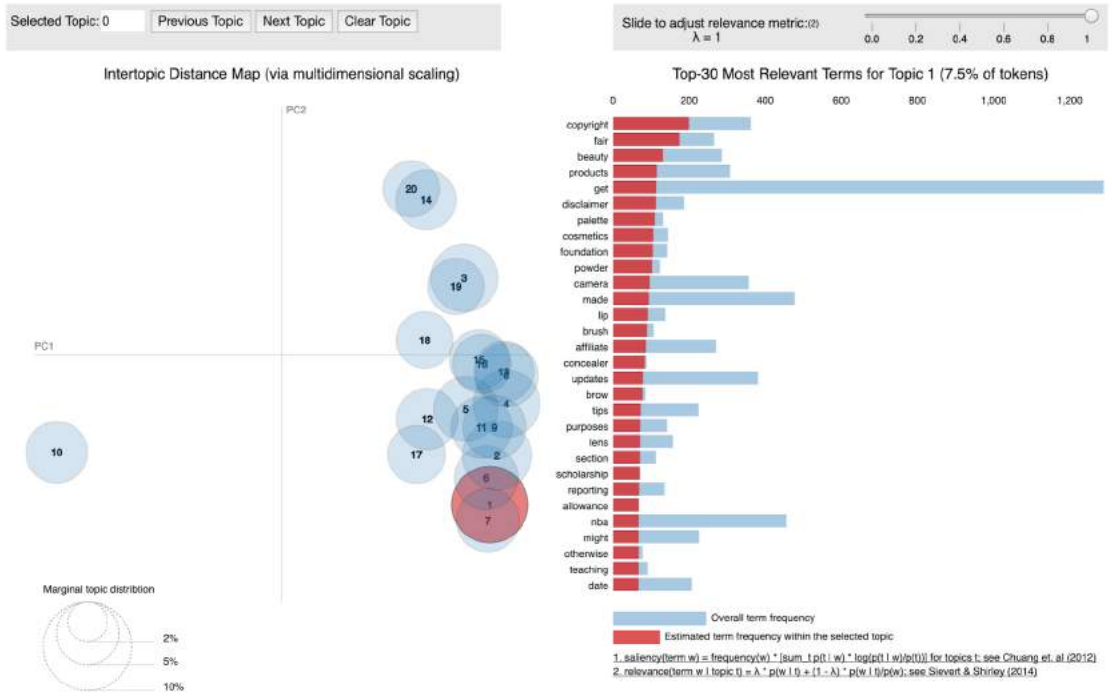
- Embedding Layer
  - Definition the vocabulary size
  - One-hot-encoding the corpus
  - Pad documents to max length
  - Train test split
  - Build the model
    - Embedding layer
    - Classification layer
  - Compile
  - Fit and Evaluate

- GloVe
  - Initialize the tokenizer object, and fit the tokenizer on the whole corpus
  - Converts each sentence into a sequence of numbers
  - Padding with the same length
  - Extract features from the pre-trained word vectors
  - Build the model
    - Embedding layer
    - Classification layer
  - Compile
  - Fit and Evaluate

Appendix: Topic Modeling



Politics



Beauty & Cosmetics