Analyzing YouTube Trends

Ananta Ramesh

WHAT

- YouTube is an online video-sharing platform which allows users to upload, view, rate, and share videos.
- There is a category called age-restricted videos which are only available to registered viewers over 18.
- Video descriptions are used to provide a viewer with more information on the channel, the video, and other similar videos.
- Hence, analyzing the sentiments of the videos would help provide more information as to what kind of videos tend to trend.

WHO WOULD BE INTERESTED

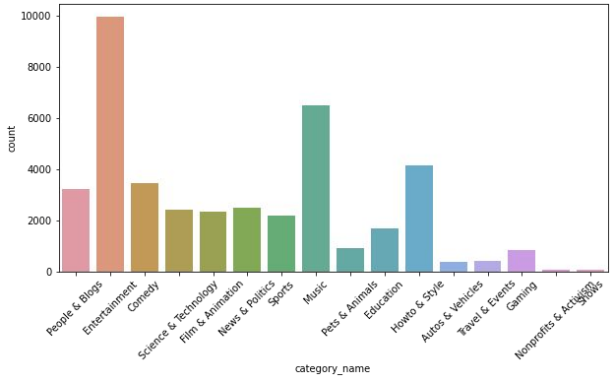
- Parents who are interested in understanding whether their child's watched content is appropriate (maybe too negative even if it is technically age appropriate?)
- Ad companies (knowing what types of videos tend to trend will help them understand what types of channels to monetize)

DATA

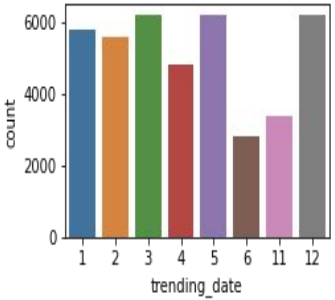
- Data on YouTube videos that trended between Nov 2017 and June 2018
- Source: Kaggle
- A series of statistics about every video, including:
  - Number of likes
  - Number of views
  - Number of comments
  - Description
  - Title
  - Category
- 6351 unique videos in the dataset
- 1 row per day of trend (hence a video can cover multiple rows)

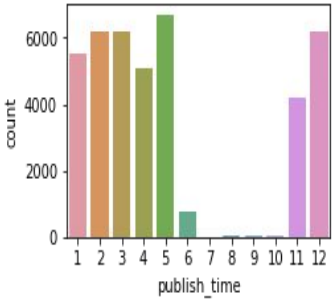
EXPLORATORY DATA ANALYSIS

Number of videos per category



Distribution of videos across trending and publish dates

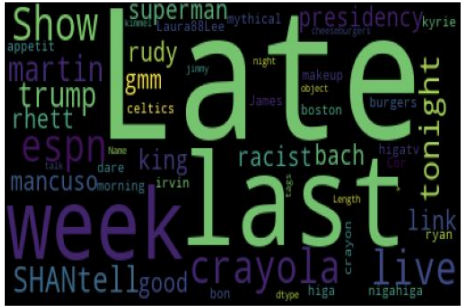




WORD CLOUD OF VIDEO TEXT

Word Cloud of Video Titles

Word Cloud of Video Tags



Word Cloud of Video Descriptions

Word Cloud of Channel Titles

METHODS FOR SENTIMENT ANALYSIS

- Sentiment Analysis:
  - understanding the context and emotion behind specific text
  - helps examine the ratio of positive to negative engagements per topic/category
- Lots of Python libraries available to perform sentiment analysis
- Two popular ones:
  - Textblob
  - NLTK
- I decided to use textblob since it has a less steep learning curve

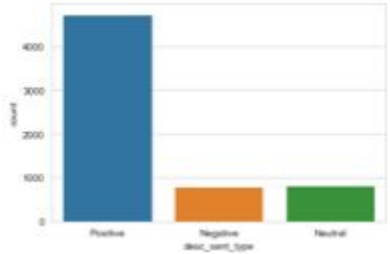
DATA CLEANING

- Overall, dataset was pretty clean.
- All numerical columns were within range - no invalid values.
- Only column with missing values - Description (which is the column I am going to employ for sentiment analysis).
- I decided to fill in the missing values with ‘No Description’ rather than deleting them since I thought it would be interesting if they contained some useful information within them.
- I cleaned the description column by using regular expressions to expand contractions, remove hyperlinks, remove non-alphabet characters, etc.

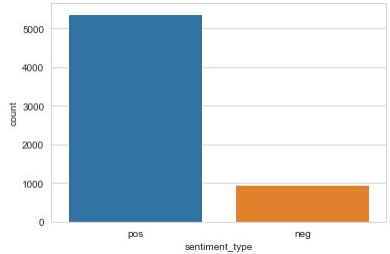
MODELING

- I used Textblob to calculate the sentiment and the polarity of each description - to understand which videos tended to be the most polar, what were the most common sentiments, etc.

With default PatternAnalyzer



With NaiveBayesAnalyzer



INSIGHTS & NEXT STEPS

- Most videos are classified (by both models) as ‘positive.’
- YouTube description data (after cleaning) does not contain a lot of useful text for accurate sentiment analysis

- These insights can be used to build a browser extension to ensure that, if kids are watching, the videos available to them are positive and not excessively polarizing
- Ad companies can also use it to better predict viewer engagement to ultimately decide what channels they should be monetizing in the future