YouTube Data NLP Analysis

Instructor: Dr. Edward Stohr

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Business Intelligence & Analytics

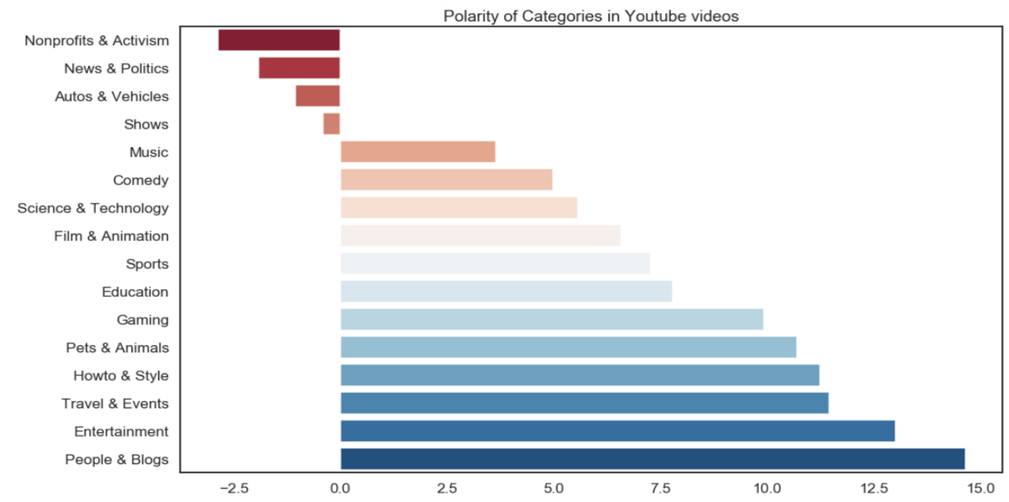
1. Objectives

- To discover the viewer's perception of videos in different categories based on analyzing the tags, description, likes and dislikes, views.
- To further develop a classification and verify if videos have been tagged in the right category.
- To explore the correlation between video publish time and probability to trend.

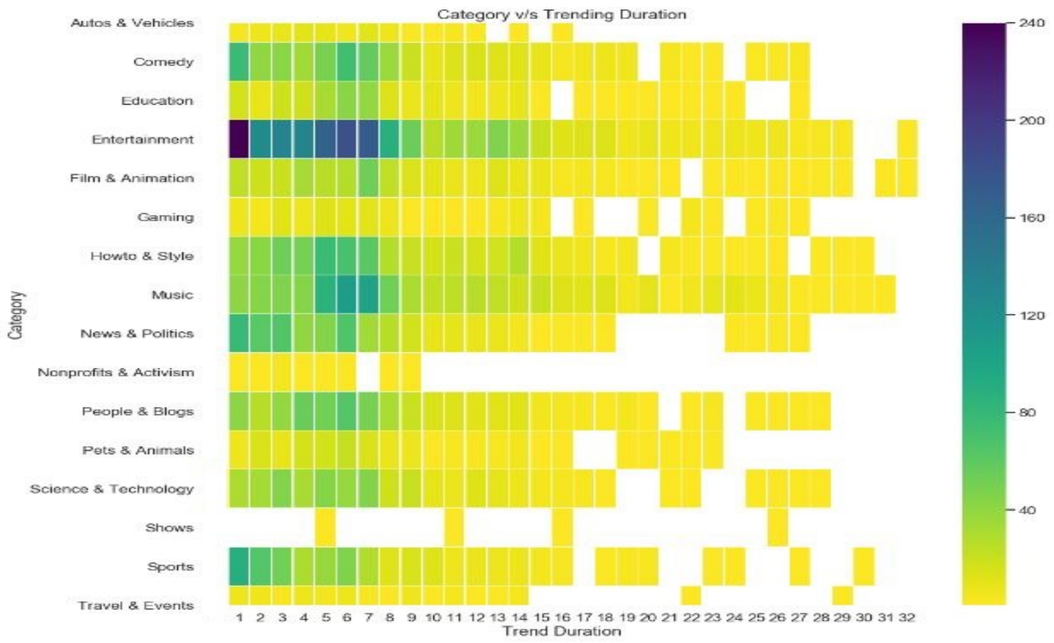
2. Recommendation

- Our team built and testified models from statistics, machine learning and deep learning aspects and improve accuracy from 81% to 97.8%.
- Our team developed a model to recognize the best publish dates and trending probability of a video. The model showed that videos published during the first 8 days of the year are more likely to trend comparing to any other time in the year.
- According to the polarity analysis result, Entertainment and Peoples & Blogs acquired the most positive sentiment score.

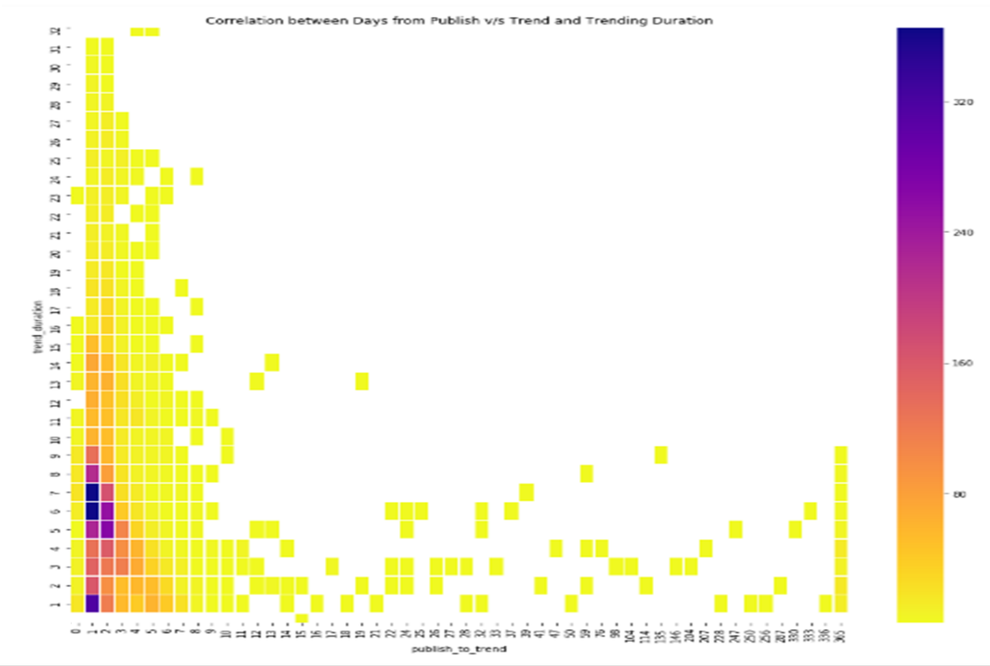
1. Polarity Analysis on tags



2. Heatmap of the trending duration of videos

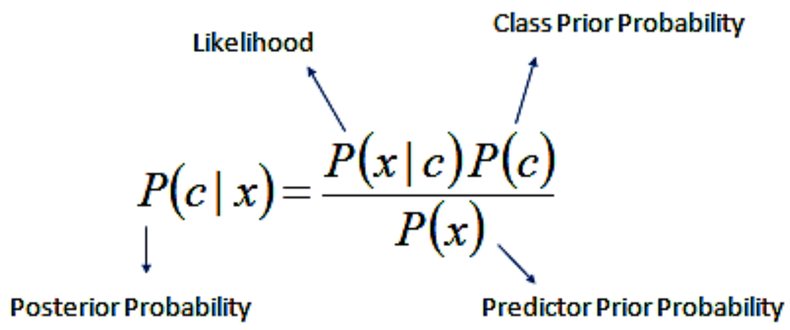


3. Heatmap of the publish time of trending videos

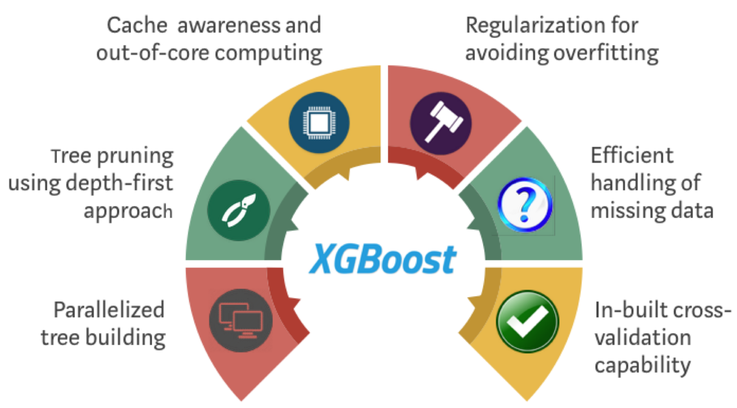


Modeling

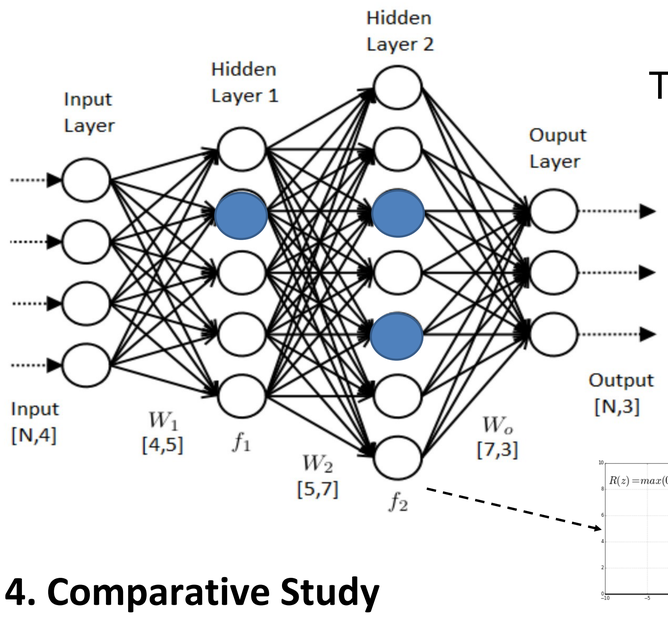
1. Multinomial Naïve Bayes (Baseline Model)



2. XG Boost



3. Artificial Neural Network (3 layers)



Two Hidden Layer: *Relu*

Output Layer: *Softmax*

Optimization:

Adam

Early Stopping

Drop out

Grid Search

4. Comparative Study

	Naïve Bayes	XGBoost	ANN
Accuracy	81%	94.3%	97.8%
Time (mins)	1.2	6.3	21