THIS IS NOT A CLICKBAIT

The Relationships Between Language and Virality on Youtube

Peter Kwak Dorian Raboy-McGowan Bocar Wade

INTRODUCTION

In this project we examine how video titles and categories of Youtube Videos influence virality. Explicitly, our input was **a sparse feature matrix of all unique words** found in titles and all **video categories.** Our output was a binary labelled data (viral or not). Our goal with this project was to **expose biases** in YouTube's algorithm for featuring videos over others.

Defining Virality

We define a video to be viral if its view count is among the top 90th percentile of all of the videos in its respective category. Each category is defined by Youtube's categorization.

Title	Views	Category	Viral
"Ed Sheeran Acoustic Live Cover Official Remix Official Video Lyrics Session"	33,523,622	Music	1
"I Tried To Plan The Perfect Wedding"	961,644	People & Blogs	1

(Note: This figure represents why videos need to be classified as virality or not with respect to video category.)

DATA AND FEATURES

Dataset: "Trending YouTube Video Statistics"
Source: Kaggle.

Data Processing Procedure

① Duplicates removed.

② Video category feature added.

③ Non-Alphanumeric characters removed.

④ Sparse feature matrix created (~30,000 features).

Title: "smart girl shows how to build candy dispenser"
Column Labels = ['cheese', 'smart', 'lego' ... 'mars', 'how', 'toronto' ...'Sports', 'People and Blogs' ...]
Features = [0, 1, 0 ... 0, 1, 0, 0, 1, ...]
Varlity = [1]

(Note: red represents the word feature section and blue represents the category feature section)

⑤ Partitioned data set into 80/20 split for logistic regression and binary classification neural network.

LOGISTIC REGRESSION

Tool: scikit learn

Logistic Regression

- Scikit learn model used (binary classification)
- Served as a baseline model to identify areas for improvement.

Process

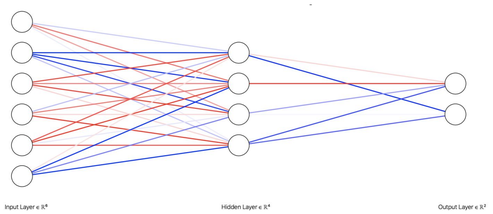
① Update weights for each feature title word and video category for thousands of training steps.

NEURAL NETWORK

Tool: Pytorch

Our Neural Net

- Binary Classification Neural Network
  (Template from Medium)
- 2-layer NN
  - Input Layer: ~30,000 Nodes
  - 1 Hidden Layer: 20,000 Nodes



Process

① Split tensors into 80-20 train-test

② Over 1000 epochs

- Predict output using forward function
  Note: with relu activation function
- Calculate Cross Entropy Loss
- Rempute weights

③ Perform a prediction on test using NN

- Bases classification on the weights of each our the output nodes

④ Perform classification report using sklearn

RESULTS

Logistic Regression

Measurement	Value
Accuracy	0.8964
F1	0.063
Precision	0.772
Recall	0.0328

(Note: These results come from running over all training examples.)

Neural Network

Due to GPU limitations we attempted to compute our neural network on with a sample size of 1000 (800 train, 200 test) and the following results were found.

Classification Report

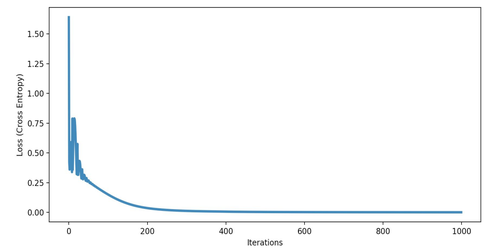
Confusion Matrix

	Actual 0	Actual 1
Predicted 0	**176**	**1**
Predicted 1	**0**	**23**

FN : "Red Velvet 레드벨벳 '피카부 (Peek-A-Boo)"

Overall Accuracy: 0.995
F1 Score: 0.995

Loss Reduction



Highest Weighted Words

1. **Ft** -> *"Am I a better chef than Gordon Ramsay? Ft. Gordon Ramsay"*
2. **Crazy** -> *"My Crazy Prom Story "*
3. **Sick** -> *"Movies That Made People Sick In The Theaters"*
4. **Love** -> *"Love Songs Piano Music 2018 - Love is not getting, but giving - Piano Instrumental Music 2018 "*
5. **125** -> *"$125 Fake iPhone X - How Bad Is It?"*

ANALYSIS

Logistic Regression

- Baseline outputs a misleading high accuracy score.
- Recall score indicates only 3% of the viral videos were classified as viral.
  - Conclusion: algorithm is simply classifying all of the data as not viral.

Neural Network

- Neural network algorithm achieves nearly 100% overall accuracy.
  - Such high accuracy indicates overfitting.
- As more iterations are implemented, the found weights result in better classifications.
- The highest weighted word, "Ft", is strongly linked to virality because it directly correlates with popular figures appearing in the respective videos.
- Words such as Sick, Crazy, or Love add shock value to video titles which seem to indicate virality.
- False negative found due to limited features available to classify it despite it having high views (>151 million).

NEXT STEPS

- We hope to utilize allocated cloud GPU credits in order to run our neural network using all 30,000 data points.
  (24,000 train/3,000 validation/3,000 test)
- Inclusion of additional features provided in our dataset to improve classification accuracy.
  (Number of Likes, Dislikes, Comments)
- With our improved implementation, we hope to find weighted words which can be more directly linked to virality.
  - With better results, we can better evaluate how word choice relates to Youtube's algorithm for videos.