CSC 422 Midterm Report: Effect of Titles on Trending YouTube Videos

Vinhson Phan, Sagnik Nayak, Rhea John
North Carolina State University

1 INTRODUCTION

We want to use the Youtube Trending data set to study what, and how different attributes of a video title affect Youtube's trending section. We will analyze various parts of the string title; attributes like capitalization, sentiment, title length, punctuation, etc. We will correlate these attributes with data such as likes, dislikes, comments, to evaluate the engagement and influences of these various title attributes in the Youtube Trending Page algorithm. Our study aims to find how different titles affect the way the Youtube algorithm behaves to certain videos and therefore how the video makes it to the trending page, which in turn directly affects the viewership of the video. Our study can give a deeper understanding of the effect of title phrases and punctuations in video performance and hence provide an insight on how creators can frame their titles for the best performance on the platform. Our learning model will help predict the likelihood of a certain video title to make it to the trending page which would be valuable feedback for content creators and Youtube themselves.

1.1 Related Works

(1) Understanding Digital Ethnography: Socio-computational Analysis of Trending YouTube Videos. [1] This paper utilizes a similar dataset from Kaggle to draw out the correlation between comment count and sentiment analysis from the comment section. It is related to our problem because we also conduct sentiment analysis, the presence of or lack of, in the title and the effect of it on views and dislikes. The paper took a similar approach to initially studying the correlation between various attributes with the comment section to narrow their research and define their approach going forward. Additionally, the paper provides a rather more statistical insight on the dataset which is useful when considering what approaches to take when training our neural network with a set of attributes.

(2) Predicting the citation counts of individual papers via a BP neural network. [4] This paper discusses the following: tuning neural networks, optimal learning rates and loss, measurement of error, performance comparison across learning models, feature creation, feature importance, and how BP NN outperforms XGBoost, RF, LR, SVR, KNN, and RNN. We are able to relate our current problem of classification that uses feature creation from a paper to the one mentioned in the paper. The paper provides good evidence in using a BP neural network in our approach as compared to other mentioned analysis methods such as KNN.

(3) Comparing Rewinding and Fine-tuning in Neural Network Pruning. [3] This paper discusses how to improve the performance of a neural network by pruning features that overfit which affect the testing accuracy of the model. It also discusses optimal activation functions for various use cases and the weight rewinding method based on the initial training performance. This paper is related to our current problem of figuring out how to successfully use activation functions with the attributes we deem fit for the neural network such as title length, sentiment analysis, presence of exclamation point, etc, and applying appropriate, adequate weights. Additionally, the paper provides guidance in our approach on how to tune the neural network through the thorough techniques described in the paper.

2 METHOD

2.1 Approach

Our study starts with the Youtube Trending Dataset on Kaggle. The dataset contains information about Youtube videos that made it to the trending page with columns such as, trending date, title, tags, and performance metrics: likes, dislikes, and comments. We drop features not relevant to our study (we are concerned with the title and performance metrics) and create features relevant to the learning model. These features include a percentage of capitalization in the title, containing fully capitalized words, exclamation, and question mark usage, and sentiment scores using the nltk library. After constructing new features from the title we will use holdout validation and split our dataset into training and testing sets. The new features created based on the title will be correlated to the performance metrics of the video. We will use a Neural Network to build our classification model. Then, we will train the Neural Network with the new features in the training set to predict the trending page classification for a video. We will tune the Neural Network based on the performance in the first few epochs to set out weights for particular features to best predict the classification.

2.2 Rationale

The Youtube video trending position depends on several factors which are weighted by some internal algorithm in the Youtube ranking system. Each video has several attributes and these attributes affect the likelihood of a video making it to the trending section of the video. We want to tune our learning model so that it can learn from the data and set weights for each attribute to mimic its importance as seen on Youtube. For this purpose, a Neural Network seemed like the best approach for this problem as they are inherently designed to learn and process multiple data attributes and predict class variables based on varying weights for these attributes. We will train our Neural Network on the features created so that it can then accordingly weigh in each attribute to best classify whether a video will rank in the trending section or not.

3 EXPERIMENT

3.1 Dataset

Our data set contains information on trending videos with attributes like title, likes, dislikes, tags, etc. We will pre-process this data set to cater to our needs of a title, ranking on the trending page, and the amount of time on the trending page. We obtained the dataset from kaggle.com and according to the documentation of the Kaggle dataset, the data was collected using the Youtube API.

3.2 Hypotheses

We predict that a title that contains a fully capitalized word is more likely to trend. Another prediction we have is that containing an exclamation mark leads to more views and in turn a higher likelihood of trending. The title length should be near 50 for the highest chance of trending. A combination of these attributes along with proper weighting should result in the classification of a trending video. These predictions come as a result of some of our data analysis as well as from our anecdotal experience with what we have seen trend.

3.3 Experimental Design

For pre-processing the data, we imported the dataset into a Pandas dataframe and dropped the columns that we felt were unrelated to the experiment. Using the title attribute we created several features. We felt that these features would be of particular significance. First, we created a feature for title length, because it is a numerical attribute that can have an optimum amount for the likelihood of trending. Next, we checked for both the existence of "!" and "?" within the title. We also checked for the existence of a fully capitalized word in the title as well as the percentage of letters that were capitalized. Lastly, we used pythons nltk library to do some sentiment analysis on each title through VADER. We chose VADER because of its availability and versatility. For this attribute, we use the compound sentiment score, where a negative value is a negative sentiment and likewise for positive. The value ranges between [-1,1]. By using VADER, which uses valence-based lexicons, we can see sentiment intensities [1]. Also, VADER was particularly good at classifying the sentiment of tweets (96% F1 score) [2], which are another form of social media, so it may have good accuracy for youtube titles. This preprocessing was done through python and applying functions to the title column of the dataframe. Each of these features was created through relatively simple functions.

We have not yet created a machine learning model for this problem and plan to do this by the end of this project. As of now, we have only done some statistical analysis on some of our attributes. We will be using likes and views as performance metrics to see to what extent a video trends. In the next section, we will discuss some findings. We do plan to use a neural network since it can take into account the weight of certain attributes using these metrics rather than just classifying based on raw probabilities. These metrics of views, likes and dislikes seem to be pretty standard in analyzing popularity. Once we train this neural network, we will see the accuracy of its predictions on what ended up on the trending page. This will be achieved by testing on our current dataset, where all videos are trending, and perhaps another dataset of random youtube videos where only some are trending.

3.4 Partial Results

For brevity, we will mostly be exploring the results of these features as related to views. Likes correlate with views, but obviously, they are not the same. In figure 1, we can see that a sentiment score of 0, or a truly neutral title sentiment dominates the trending page of Youtube. A neural network should take into account that most trending videos have a 0 sentiment score. However, in figure 2, we can see that frequency does not account for views. Some of the most viewed videos have sentiment, especially in the slightly negative area of the graph.

In figure 3, we see that there is an optimum title length for the number of views at about 50. Also, the densest area of the graph in the middle signifies a high number of videos around that title length.

In figure 4, we see that the percentage of capital letters has an optimum amount of around 0.1 to 0.2. This seems like it would be pretty typical of any title. On the other hand, we can see that some trending videos still exist around high percentages.

Next, we have some additional stats that will not be graphed that are likely still of some significance. About 44% of videos have a fully capitalized word in their title. This is not common in regular writing/grammar so the existence of a fully capitalized word may have some bearing in what makes a video trend. Of the titles that satisfy this condition the median amount of views is 687,902 versus 677,138 for those that do not (we use the median here to account for outliers). Only 11.9% of the videos contain an exclamation mark, but there is a large margin between median view count for those that do and do not. These medians are 881,807 and 656,472 respectively. This may point towards a title with an exclamation mark performing better than a title without one. About 5.9% of the videos contain a question mark. The median view count for those who do is 619,611 versus 685,579 for those that do not. So contrary to the exclamation point, a question mark in the title may cause a video to perform more poorly.

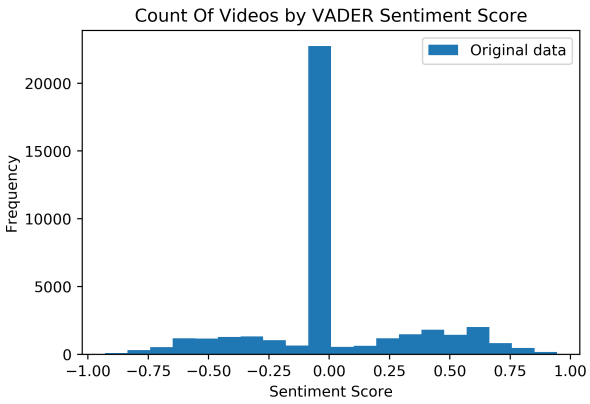

Figure 1: Sentiment Score Frequency

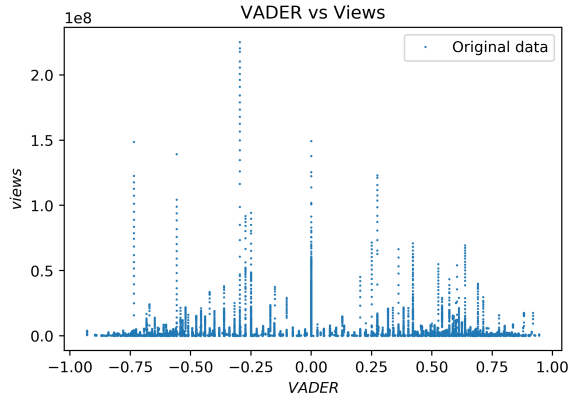

Figure 2: Sentiment Score versus Views

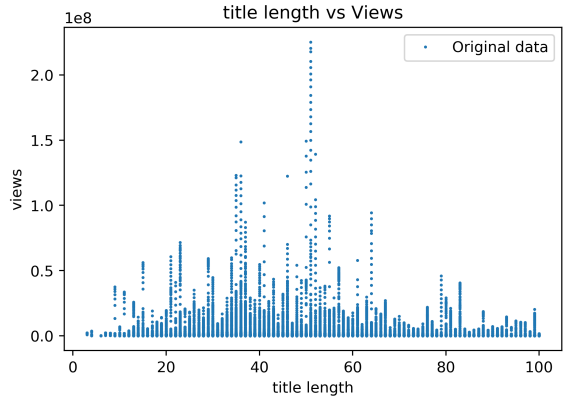

Figure 3: Title length versus Views

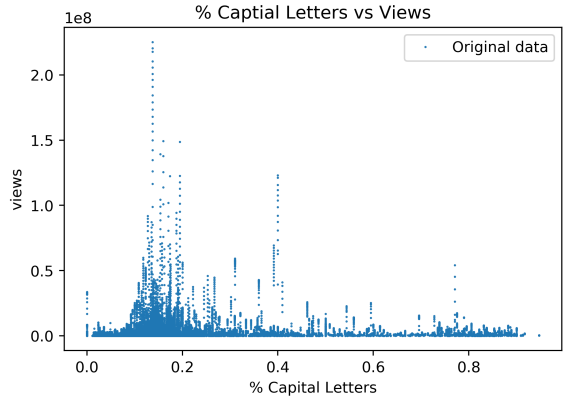

Figure 4: Percentage of Capital Letters in Title versus Views

4 PROPOSED WORK

4.1 Planned Experimental Design

We will need to shape our data attributes into a numerical form that will help the NN model best perform, we may use some form of normalization to do so. The NN will be built using libraries like tensorFlow and Keras and we will tune it with different numbers of layers to minimize our testing error. We will conduct a trade study of the best available activation and loss functions that are relevant to our project; we will start with relu and the sigmoid function to get feel for the model's performance and as we learn more, we'll implement the NN with other more appropriate activation's and loss functions like softmax. Our main challenge going forward is preparing the data for the Neural Network and then tuning the model for best performance.

4.2 Plan of Activities

Activity	Role
Normalize data attributes for NN model	Vinhson
Learn about tensorFlow + Keras libraries	All
Research trade study on best activation and loss functions	All
Train the neural network	Sagnik
Test the neural network	Rhea
Complete visualization proponents for the paper in order to display our results	Rhea
Research related mathematical formulas to provide support for our findings	Vinhson

This is our proposed plan of work, but it is subject to change. For our meetings we are planning:

- April 13, 2021: Discuss findings from research and begin normalizing data attributes for the NN. Document our experimental methods and results along the way.
- April 15, 2021: Begin tuning the NN and figuring out how to split the dataset
- April 20, 2021: Additional tuning and working on the NN, spot out any findings + work on related visualizations
- April 22, 2021: Test the neural network more and complete data visualization
- April 27, 2021: Finish up tasks and write up the paper
- April 29, 2021: Wrap up write-up for the project paper and peer review with the TA's.

5 APPENDIX

Project Github:
https://github.com/iSagnik/Effect-of-Titles-on-Trending-YouTube-Videos

REFERENCES

[1] Muhammad Nihal Hussain, Serpil Tokdemir, Samer Al-khateeb, Kiran Kumar Bandeli, and Nitin Agarwal. 2018. Understanding digital ethnography: socio-computational analysis of trending YouTube videos. In *The Eight International Conference on Social Media Technologies, Communication, and Informatics.*

[2] C.J. Hutto and Eric Gilbert. 2015. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.*

[3] Alex Renda, Jonathan Frankle, and Michael Carbin. 2020. Comparing rewinding and fine-tuning in neural network pruning. *arXiv preprint arXiv:2003.02389* (2020).

[4] Xuanmin Ruan, Yuanyang Zhu, Jiang Li, and Ying Cheng. 2020. Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics* 14, 3 (2020), 101039.