BIA 650-A

Final Project

YouTube Data NLP Analysis

Team 1

Contents

[1. Introduction 3](#_Toc26960978)

[1.1 Project Description 3](#_Toc26960979)

[1.2 Object 4](#_Toc26960981)

[2. Exploratory Data Analysis (Presentation) 5](#_Toc26960982)

[2.1 Data Descriptions 5](#_Toc26960983)

[2.2 Sentiment Analysis 7](#_Toc26960984)

[2.3 Data Imbalance 8](#_Toc26960985)

[3. Nature Language Process (Predication) 9](#_Toc26960986)

[3.1 Data Cleaning Process 9](#_Toc26960987)

[3.2 Corpus, Tokenize, Stop-words. 9](#_Toc26960988)

[3.3 TF-IDF 11](#_Toc26960989)

[3.4 Naïve Bayes 11](#_Toc26960990)

[4. Advanced Algorithm (Optimization) 15](#_Toc26960991)

[4.1 SMOTE 15](#_Toc26960992)

[4.2 Artificial Neural Network Model 17](#_Toc26960993)

[4.3 XGBoost 19](#_Toc26960994)

[4. Discussion & Limitations 20](#_Toc26960995)

[5. Conclusion 21](#_Toc26960996)

# Introduction

## Project Description

We got this data set from Kaggle (<https://www.kaggle.com/datasnaek/youtube-new>). This dataset contains a list of the top trending video on YouTube platform.  [According to Variety magazine](http://variety.com/2017/digital/news/youtube-2017-top-trending-videos-music-videos-1202631416/), “To determine the year’s top-trending videos, YouTube uses a combination of factors including measuring users’ interactions (number of views, shares, comments and likes). Note that they’re not the most-viewed videos overall for the calendar year”. Top performers on the YouTube trending list are music videos (such as the famously virile “Gangam Style”), celebrity and/or reality TV performances, and the random dude-with-a-camera viral videos that YouTube is well known for.

This dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), with up to 200 listed trending videos per day. In our project, we only use dataset from USA.

Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.

The data also includes a category\_id field, which varies between regions. To retrieve the categories for a specific video, find it in the associated JSON. One such file is included for each of the five regions in the dataset.

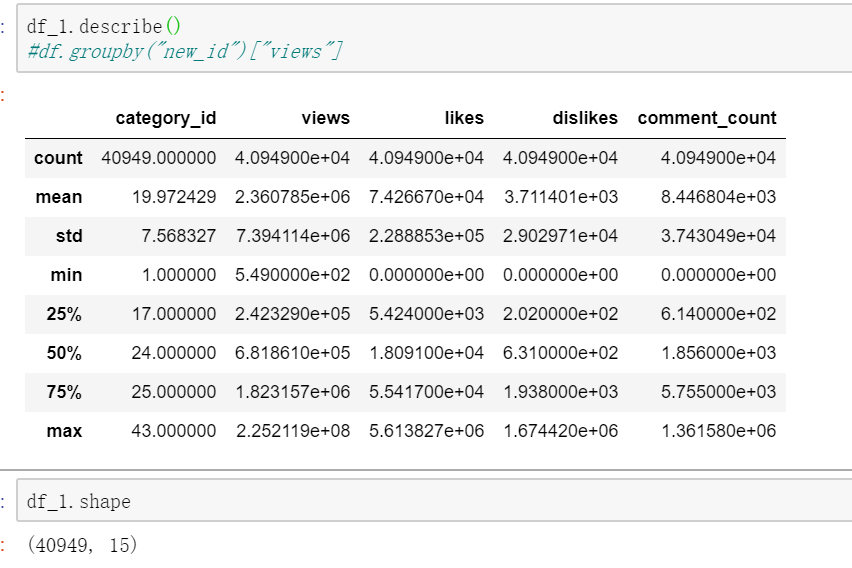
## Object

* 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.

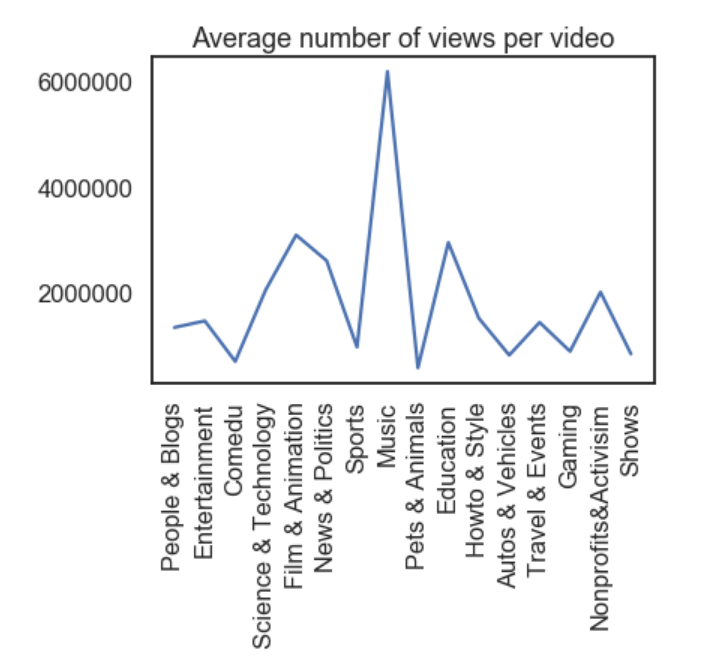
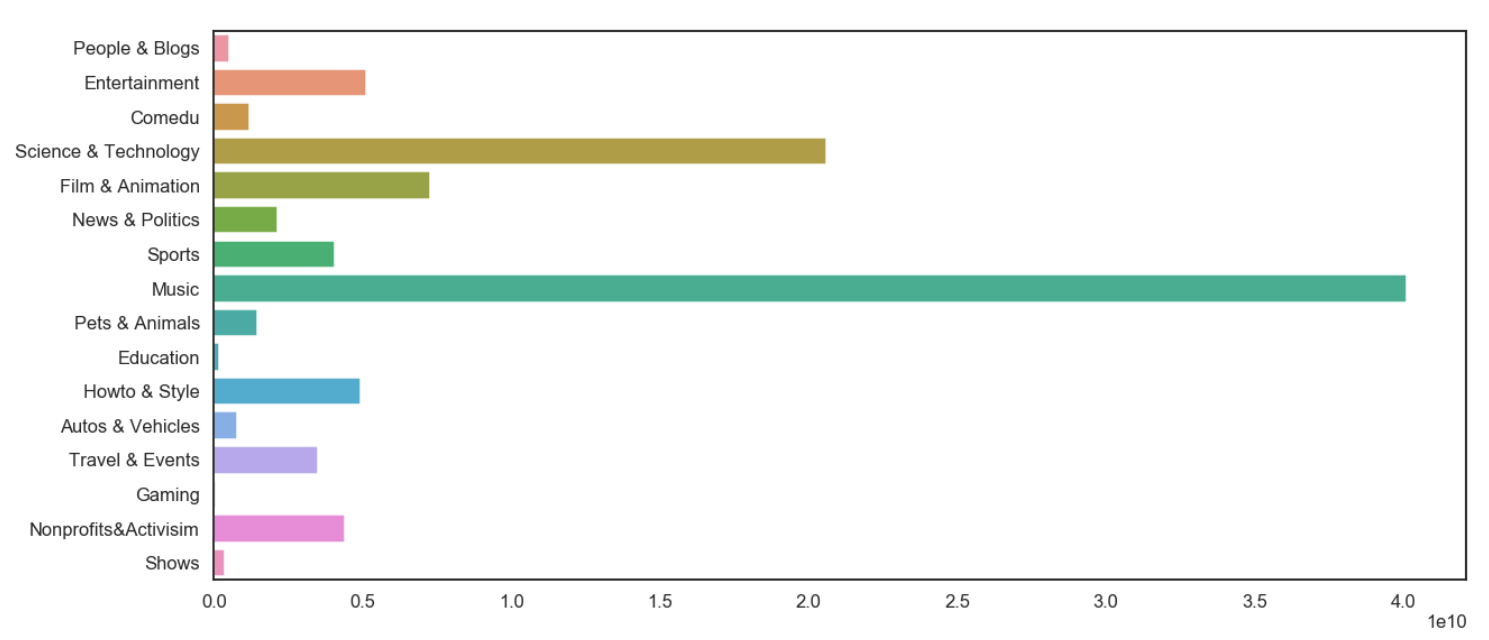
# Exploratory Data Analysis (Presentation)

## Data Descriptions

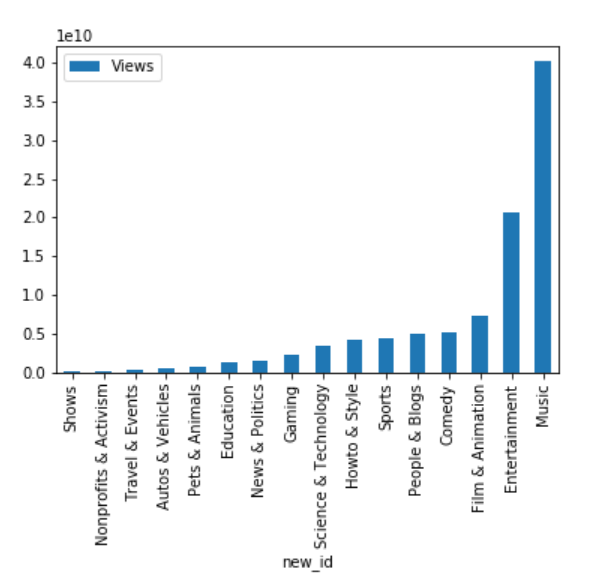
Due to language limitation, we can only process English language, so we only pick USVideo.csv as our dataset. This data set has 40949 rows and 15 columns. We have 5 columns are numeric and others are characters.

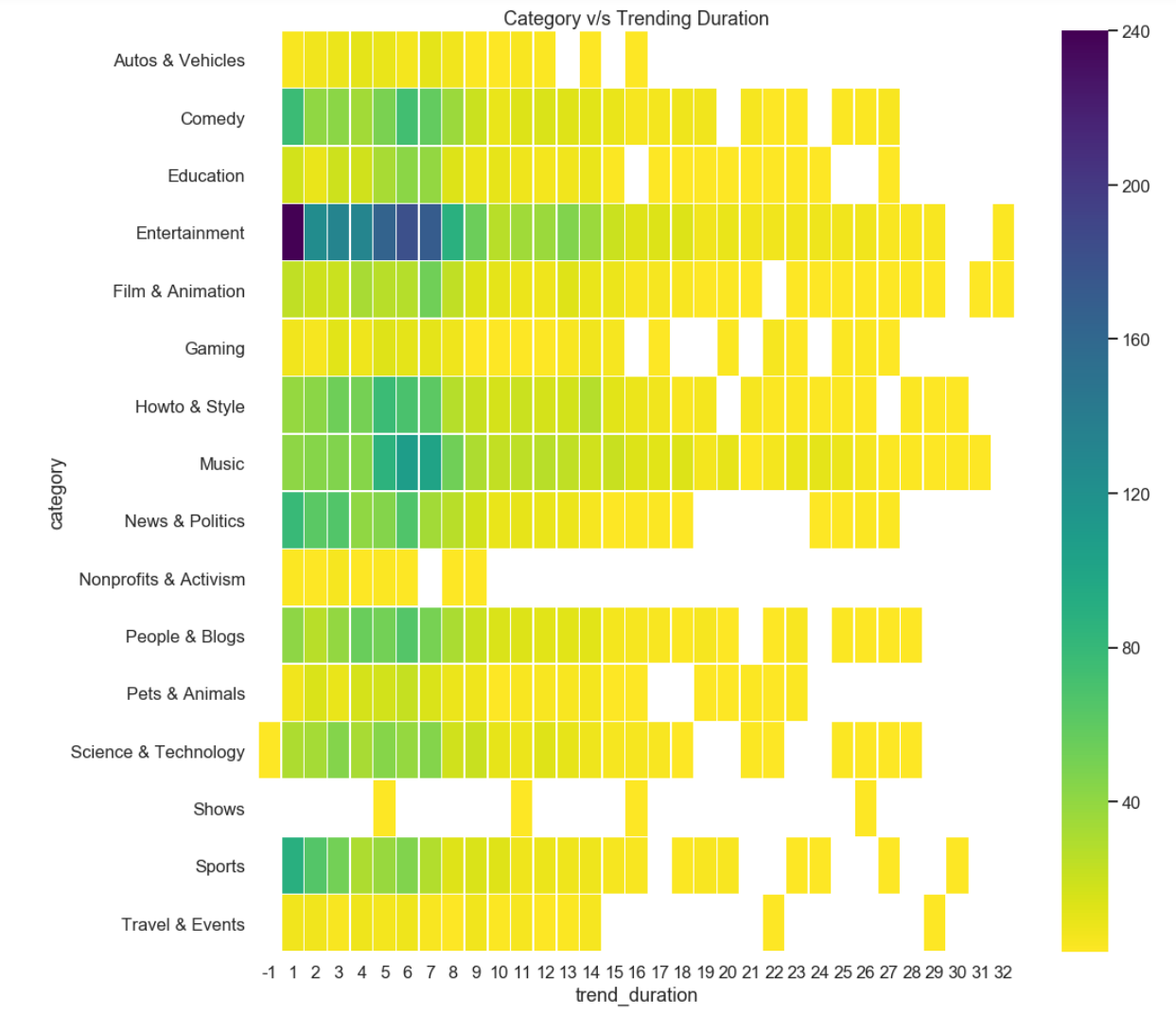


We intend to explore the relationship between category and other columns, so we try to find impact between category and categories.



According to the relationship between column ‘views’ and ‘category’, we find the imbalance between different videos. Most of people prefer watching music and entertainment context on YouTube. Other videos types, like Science, Travel and gaming, cannot attract people to pay attention even they have same number level with music and entertainment.

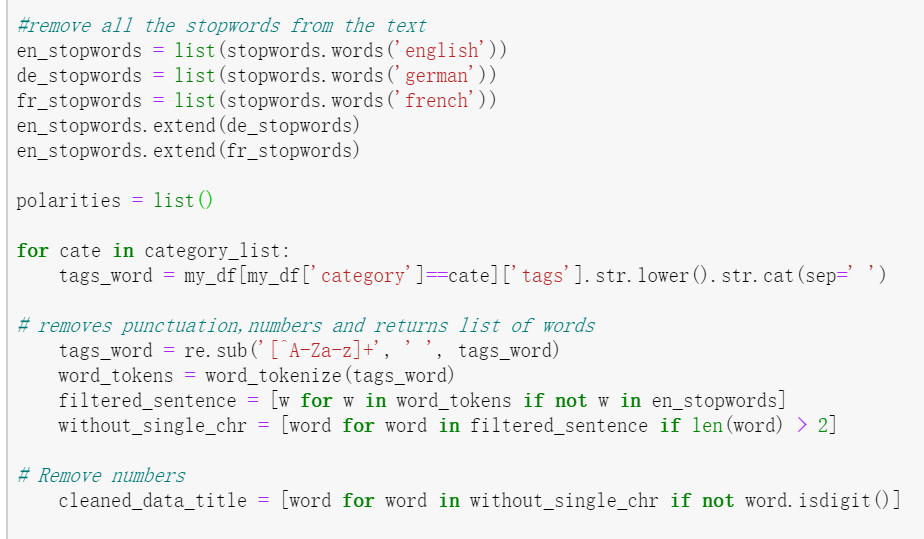


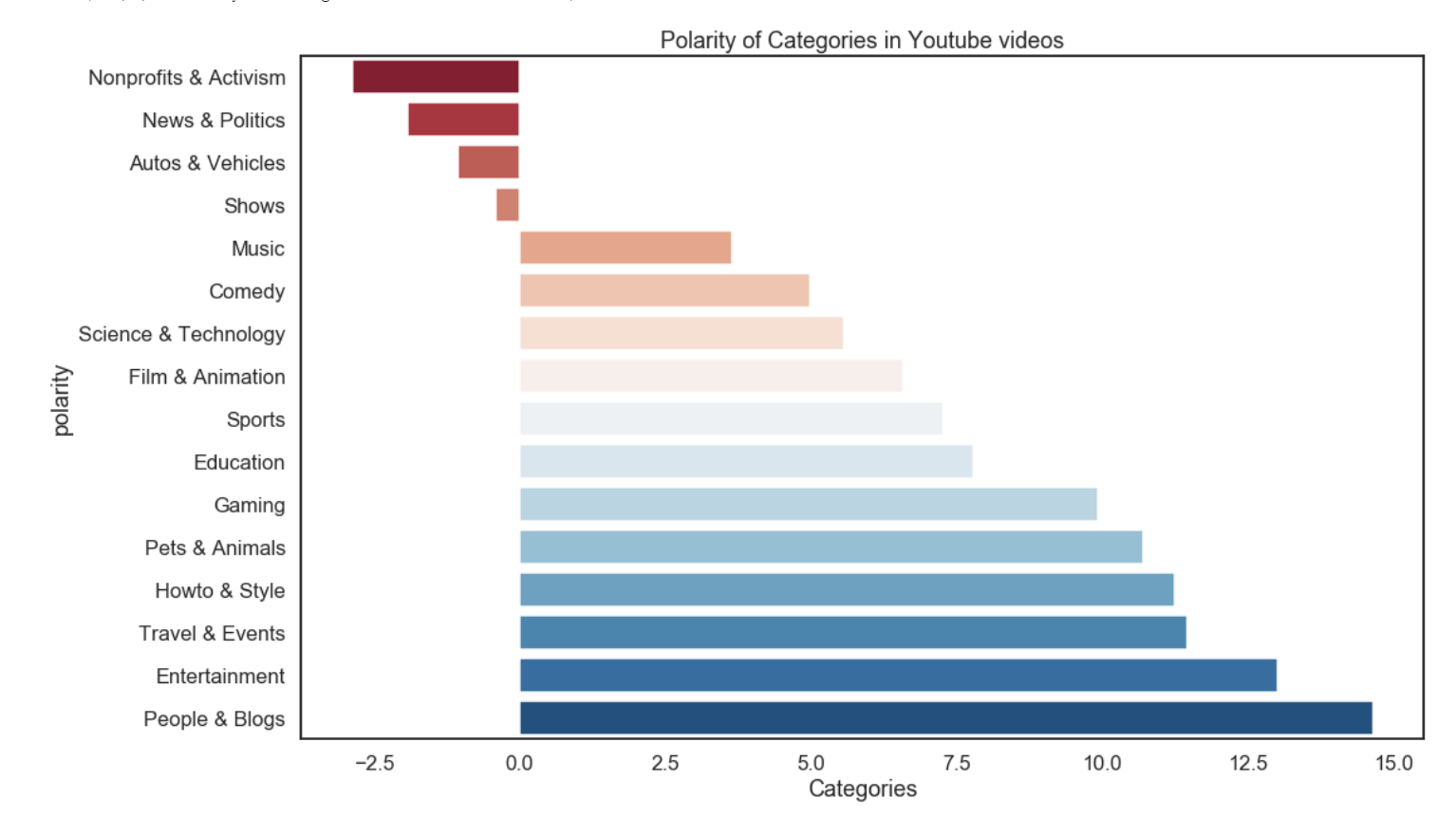


## Sentiment Analysis

We also use positive words and negative words to complete sentiment analysis.

To get an available corpus, we should tokenize dataset first, including removing punctuation, numbers, Stop\_words and special characters. We only left lower case words for analysis.





From the result, we can analyze that the words using in Entertainment, People Blogs are most positive words. Maybe they are all encourage or support words for someone or maybe just compliments for funny video.

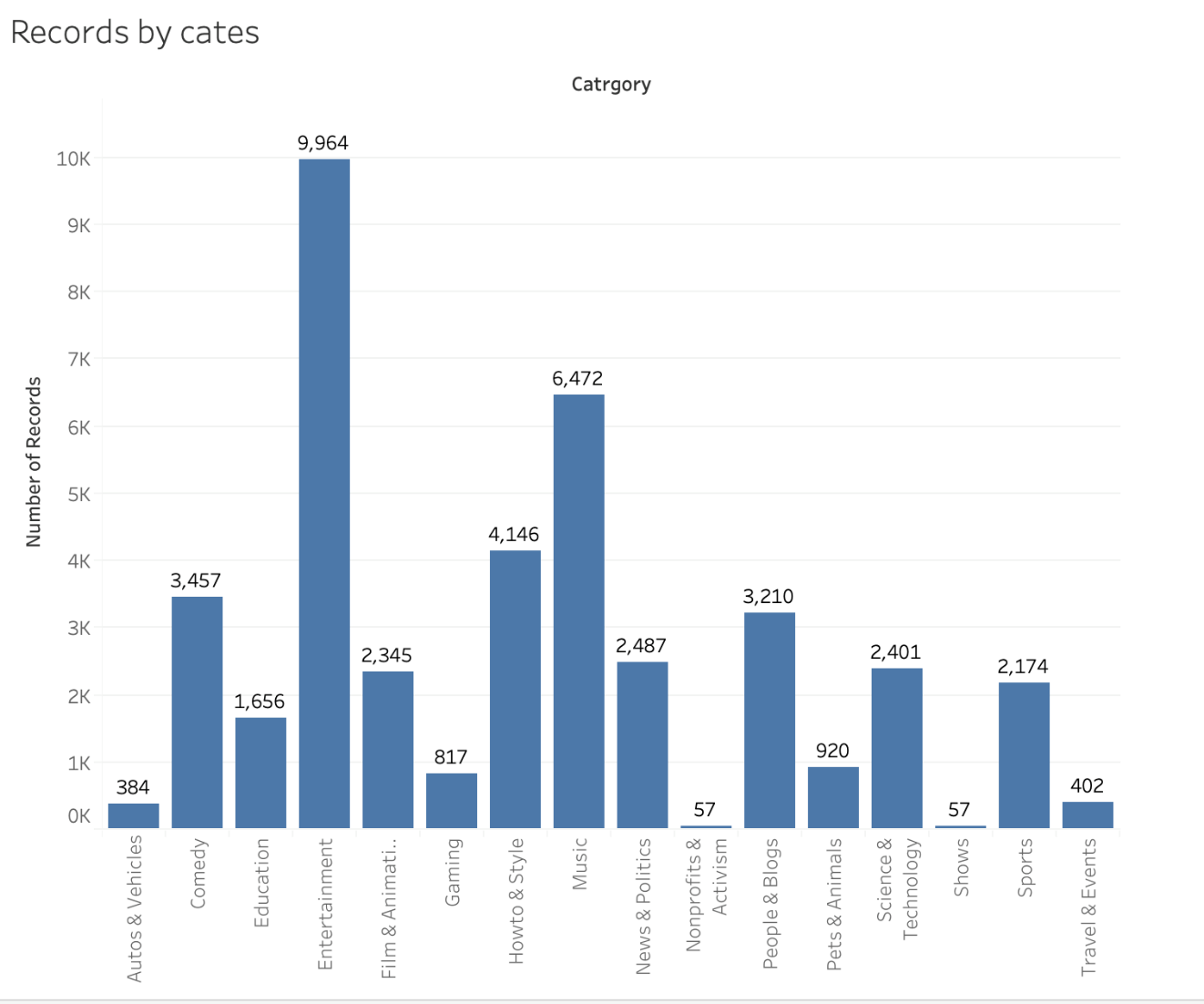
On the other side, people are mean for News and Politics. These words maybe contain criticism and censure for some politicians or international issues.

It’s a little unexpected that music and comedy didn’t present great positive trending.

## Data Imbalance

According to previous data exploration, we can find that there actually have some kind of connection between categories and other features. However, another noteworthy problem have come into sight----Data imbalance.

Some of categories of this dataset is far more data points than other categories.

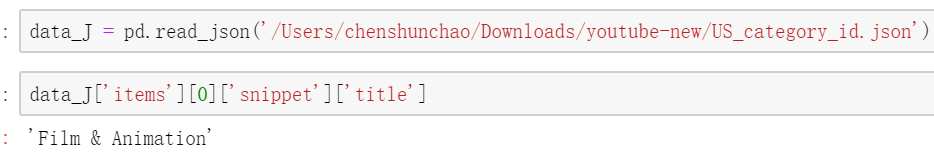


This phenomenon is easy to understand: people are more willing to see Entertainment and Music video on YouTube just for quick relax. They don’t want to learn something or think about more sophisticated international issues at their leisure. Even they maybe want to play video game themselves instead of watching online game video or show.

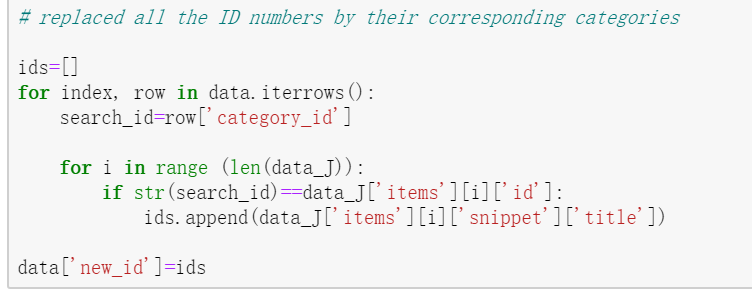
# Nature Language Process (Predication)

## Data Cleaning Process

In the original dataset, category is represented by numbers, so we should use JSON files to link category numbers with JSON category names.



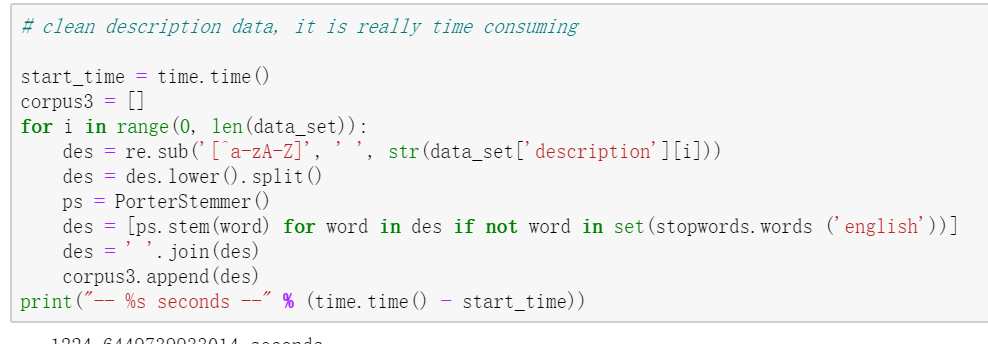
JSON files give us 32 different categories. Well actually this dataset only contains 16 categories. It’s very easy to understand, this dataset is only the YouTube trending dataset. Some of unpopular videos will not present in this dataset.



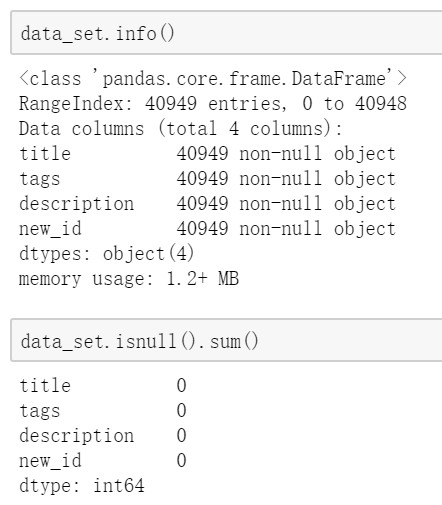
Now we have a new column to represent category with English words.

## Corpus, Tokenize, Stop-words.

In this step, we should clean all useless characters and punctuations and only keep lower case English words for us to analysis.

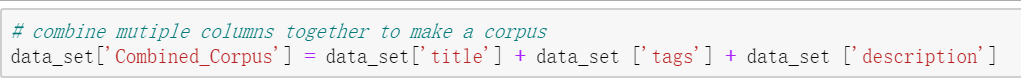


After this preprocessing, we got a three-columns (‘title’, ‘tags’, ‘description’) dataset without null value.



The standard NLP procedure including croups, tokenize, stemming, Stop\_words and preprocessing.

We only care about how ‘title’, ‘tags’ and ‘description’ will impact our ‘category’ result. We believe some words in ‘title’, ‘tags’ and ‘description’ will mostly determine category of this video.



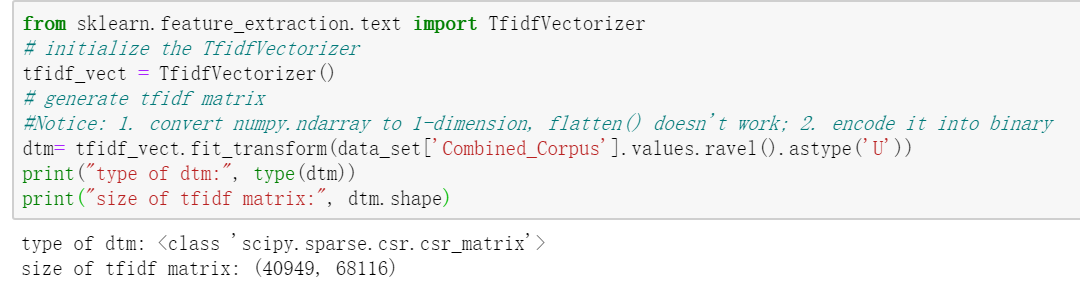
Finally, we gather all words of these columns into one group.



## TF-IDF

TF-IDF, short for term frequency-inverse document frequency is numerical statistic that is intended to reflect how important a word is to a [document](https://en.wikipedia.org/wiki/Document) in a collection or [corpus](https://en.wikipedia.org/wiki/Text_corpus).

TF-IDF can process transform characters into vectors.

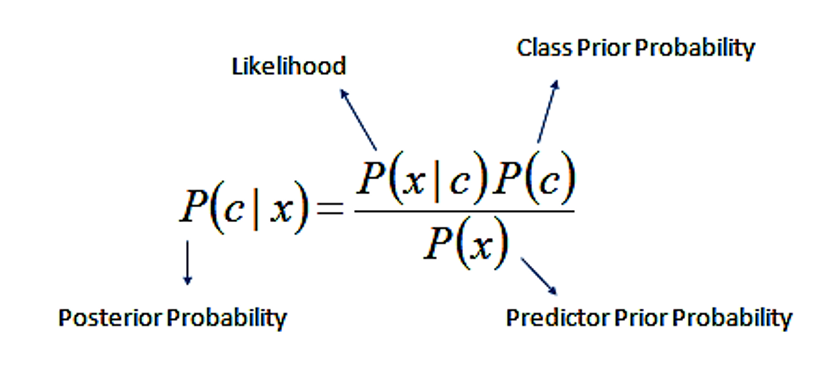


After that, we get a 40949 rows and 68116 columns matrix contains vector numbers of our corpus.

## Naïve Bayes

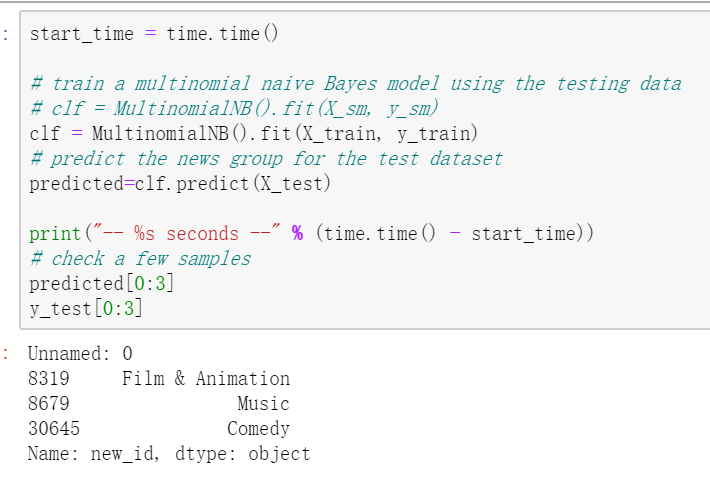
This is our first and primary step for predicting data relationship between category and other columns text information. We didn’t expect how this process will be great success. It’s just to verify our speculation.

To find out the relationship of words in dataset columns and categories. We set our corpus, which contained all the stem words from ‘tags’, ‘title’ and ‘description’, as X independent variables. And we set categories from data[new\_id] as our Y dependent variables. So, we try to find patterns between corpus and categories.

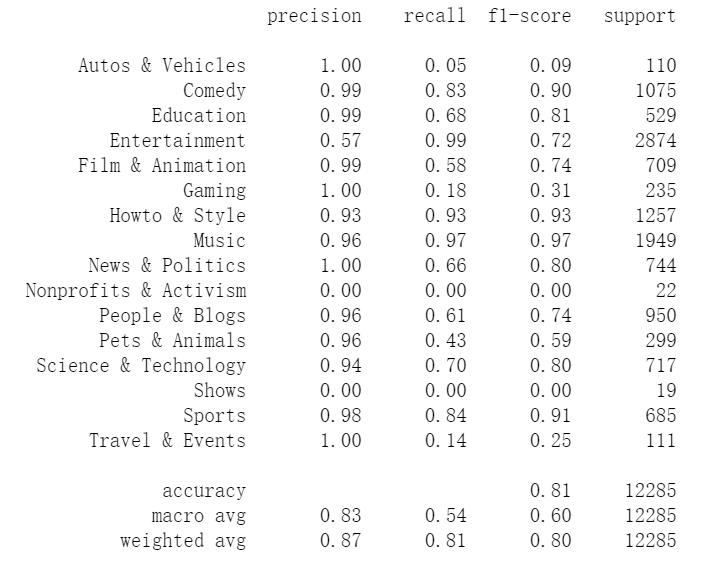


In Naïve Bayes algorithm, X independent variables---the corpus---is ‘c’ parameter in formula, the classes. Y dependent variable---the categories---is ‘x’ parameter in formula. We are actually calculating the condition probability of the corpus on categories conditions.

For example, we will calculate the probability of word ‘beautiful’ under condition category ‘Music’. If this probability is large compared to other results, we may consider this combination have some relationship.



For these categories, we get different results. We use:to evaluate our results from different models. Because Naïve Bayes is basic models, this result maybe not that correct, we will use other models to improve accuracy.

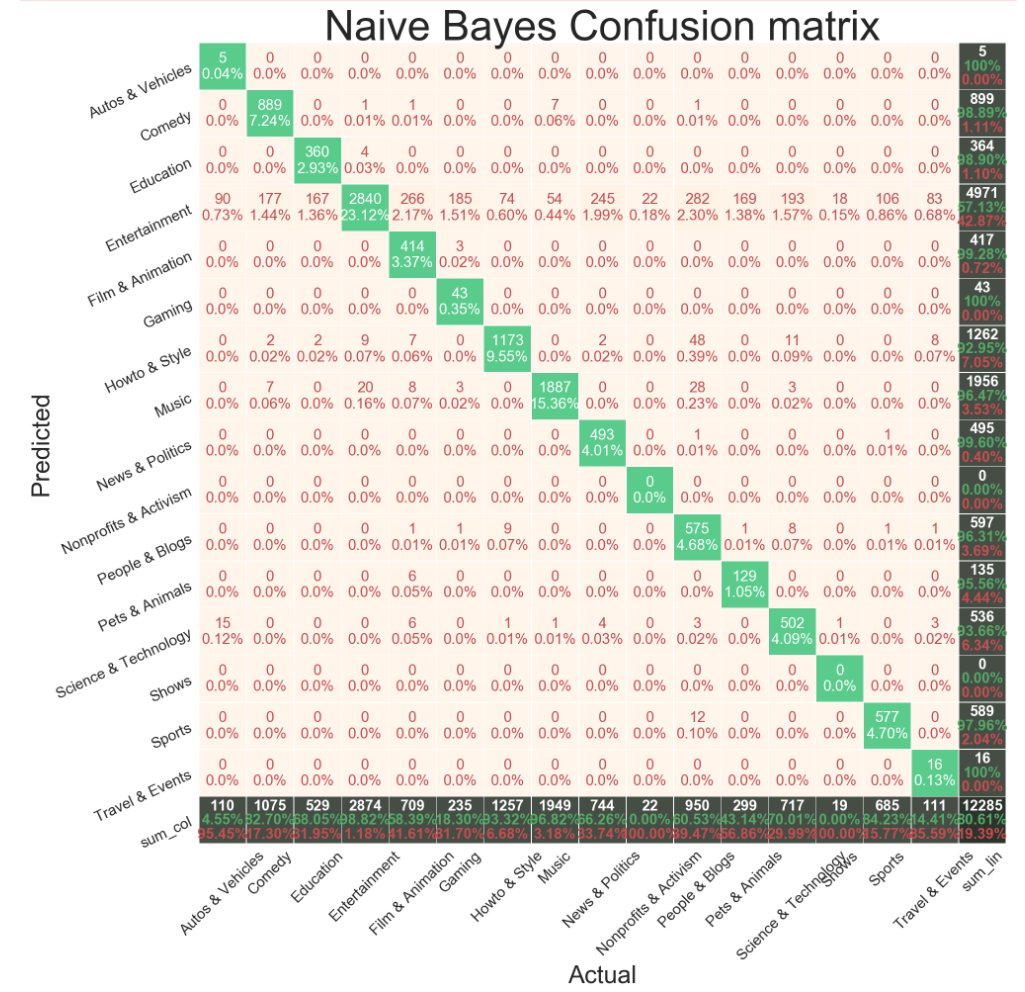


In this result, we can see that most of category have 90% greater accuracy, which means some words in ‘title’, ‘tags’ and ‘description’ can really represent this type of category.

‘Entertainment’ maybe has more common words with other categories. So, this 57% accuracy is reasonable. But other categories are performed great so far.

In this graph, if X-axis and Y-axis have the same name, their predict words and actual words will match. The number represent match numbers and percentage present how accuracy.

Well, we actually have data unbalance. The reason that some data have great accuracy might cause by their relative fewer data sample.



# Advanced Algorithm (Optimization)

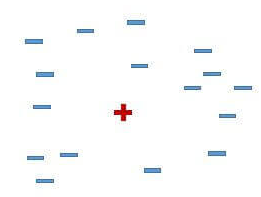
## SMOTE

In order to improve our model accuracy, we need to solve data imbalance first.

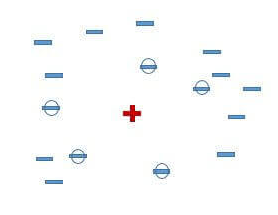
We introduce SMOTE (Synthetic Minority Over-sampling Technique) in our preprocessing data procedure. This method is going to solve the problem that some categories have relative smaller data point. For instance, in former explore, we find that ‘Entertainment’ have 4971 data point, the reason maybe cause by people like to make more words description in this category compared to other category and It’s more difficult for accurate description and tags for entertainment due to its vocabulary coverage. In the meantime, for instance, sports only have 589 data points.

So we have to find ways to increase sports data points to compare with Entertainment category.

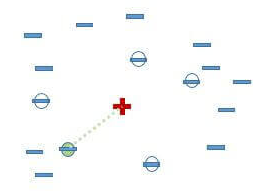
For SOMTE method, first we choose a random sample. (Suppose this sample is ‘minority’)



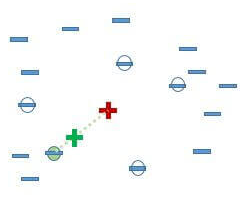
Second, we find five neighbors (k=5) of this random point



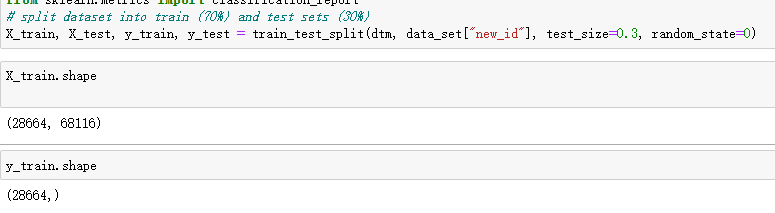
Third, we choose a neighbor simple randomly

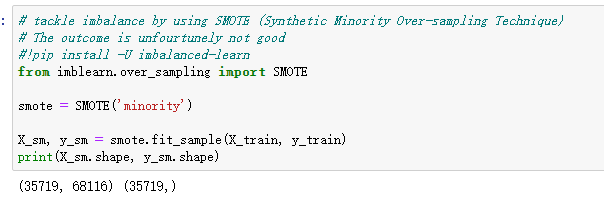


Fourth, find a point randomly on this line and this new point is our aritifial new data sample



In our model, we can see the data point increase due to application of SOMTE method





First graph is coming from our first model, we have 28664 rows, and 68116 columns, which generated from corpus contained ‘title’, ‘tags’ and ‘description’. Y variable, which have 40949 total rows, have 28664 for its 70% present for training.

Second graph is coming from SMOTE method. After this procedure, we have TF-IDF matrix that have 35719 rows and 68116 columns. And Y variable increase from 28664 to 35719, increased 24.61%. Well most increase is coming from ‘minority’ category.

## Artificial Neural Network Model

Naïve Bayes is a great NLP analysis model and we got 84% accuracy. But for a category selection project, we should improve our accuracy to 95% above. For that level accuracy, our model will have business value and can used a business demo.

So, we choose more advanced algorithm – Artificial Neural Network Model.



We can transfer table through layer to Sequential core function to build a Sequential Model.

We set our batch size to 512 and input shape to 68116. Only the first layer we need input dataset shape size and other layers can calculate shape first layer shape size.



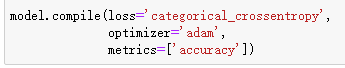
In Activation layer, we can use Rectified Linear Unit



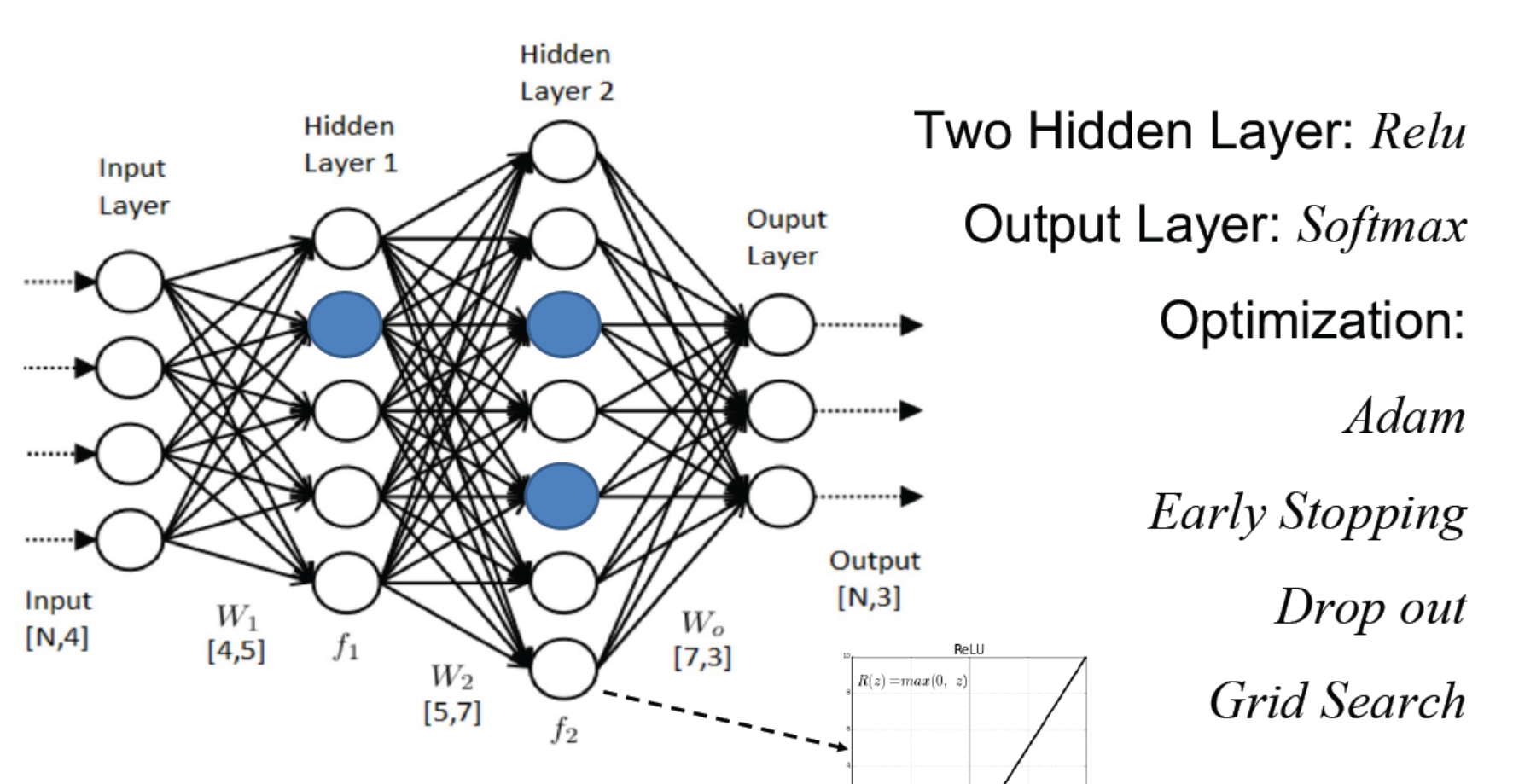
Dropout consists in randomly setting a fraction rate of input units to 0 at each update during training time, which helps prevent overfitting.

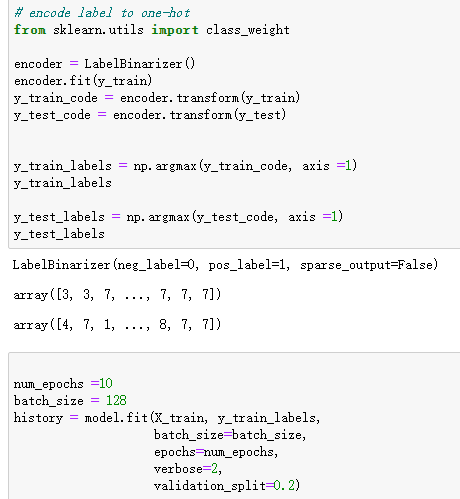


Complied our model and give lost function, optimization function and metric function.



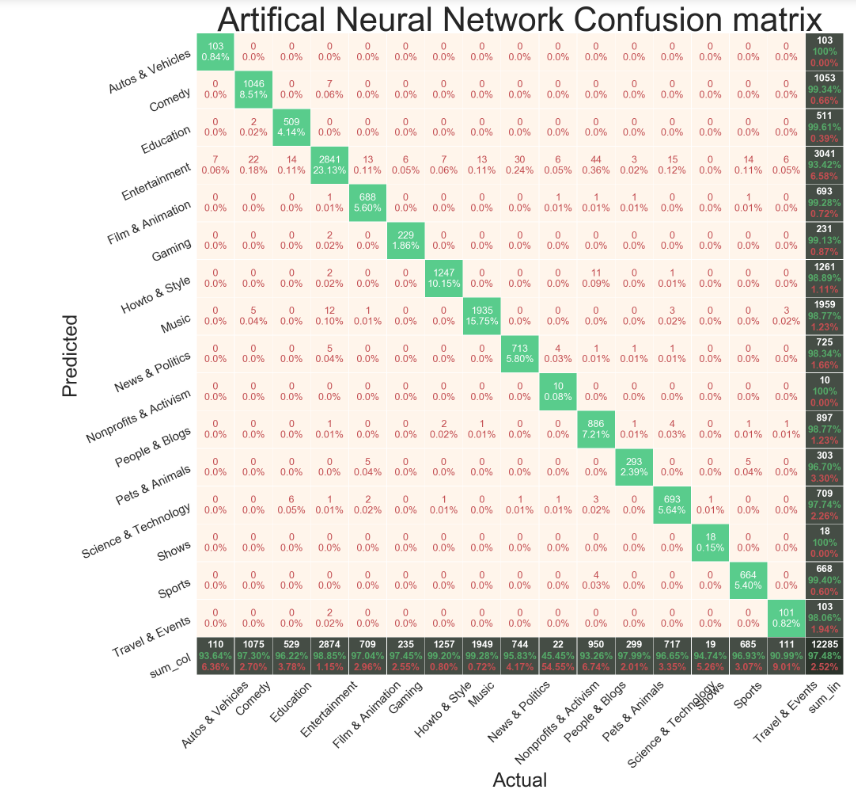
From here, we have established a three-layers model to process our dataset.



First two hidden layers, we use Relu, Rectified Linear Unit, to preprocess data. And last third layer we use Softmax to optimization our results and also third layer are output layer.

Since we have model, we can use this model to train our dataset. We set ten times and you can ses the result is become better and better through this process. For the first rotation, the accuracy improved from 80% to 93%. After three times rotation, the accuracy has been improved to 97%. Finally, after ten times rotation, our ANN model accuracy reached to 98.14%. For now, we can see we have made a success category filter model.



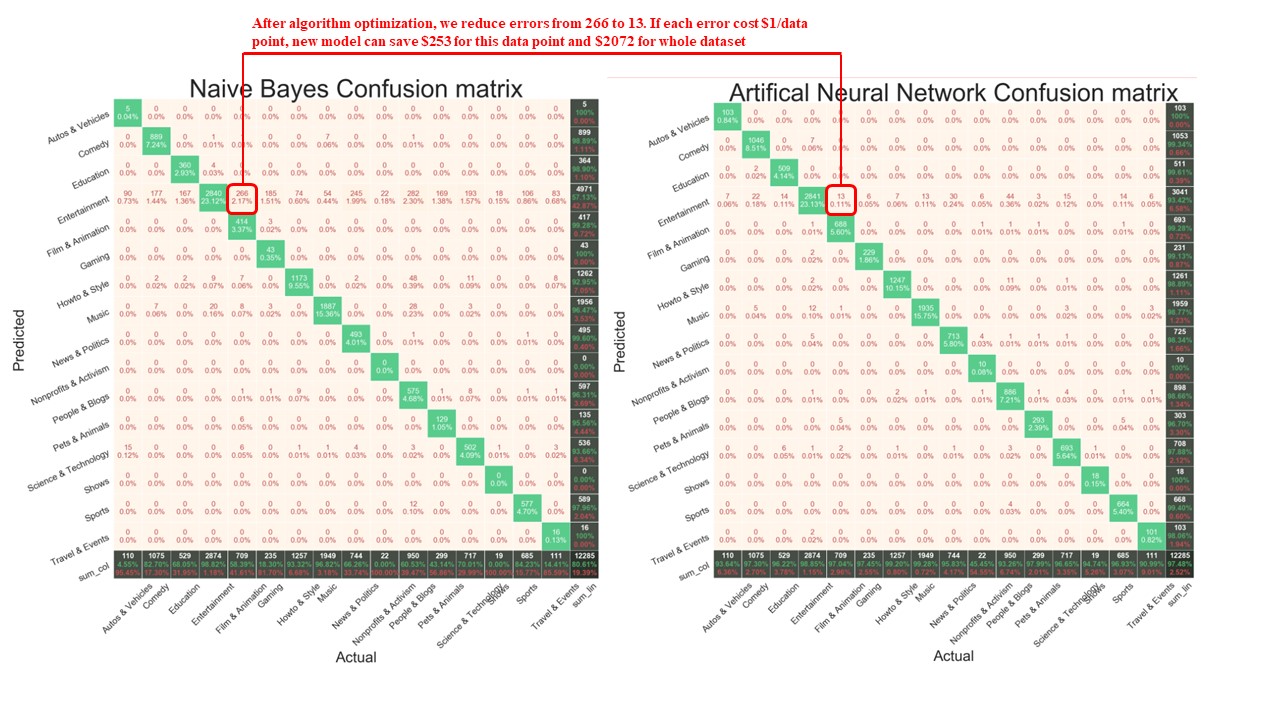
We can see from left side summary column, most of the accuracy are above 98%. Even the “Entertainment” category, which only have 55% accuracy category accuracy, have improved to 93% accuracy. That is a great improvement.

## XGBoost

Except ANN model, our team also attempt XGBoost to improve result. Using this model we can get 97% accuracy.

# Discussion & Limitations

Our model aims to server initial starter video company that don’t have algorithm to category uploaded videos and choose to use human force to label its new uploaded videos. For this type of company, we can provide our complete model for free. According to some data, one data correction by human will cost $1. For instance, in following picture demonstration, we improved our model for 266 errors to only 13 error. In this singularity modification, we can help starter company save $253. For entail dataset, we can save $2072 for 40000 observations challenge.



Our limitations are model overfittings. It’s only fitting trending videos from YouTube. For other data example, we have to re-train our model and we will face more problem to improve accuracy.

# Conclusion

For this Final project, we preprocessed 45,000+ texting data as well as numerical data and built a successful NLP classification model to category trending videos into correct field. After the tight team working, we managed to implement a complete presentation-predication-optimization data modeling procedure and finally we achieve a 98% accuracy model which can successfully predict not only the videos belonging to the major categories but also those belonging to the minor categories.