2022 International Conference on Big Data, Information and Computer Network (BDICN) | 978-1-6654-8476-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/BDICN55575.2022.00055

Youtube Data Analysis using Linear Regression and Neural Network

Jiali Fan[1*†]
[1]College of Art and Science
New York University
New York, U.S.
*jf3997@nyu.edu

Tingru Lian[2*†]
[2]College of Engineering
Boston University
Boston, U.S.
*tinalian@bu.edu
†These authors contributed equally.

Abstract—**The popularity of YouTube provides an effective way to propagate epidemic prevention knowledge by analyzing the video preferences of viewers from different locations. However, it is challenging to analyze video preferences due to the dispersed geographical locations of the YouTube viewers and the indistinguishable video categories and subcategories. This paper combines linear regression and neural networks to unravel both geographical and categorical difficulties and improve the accuracy of task-solving models. First, the YouTube dataset and extract variables are preprocessed, including categories, subcategories, countries, number of subscribers, and view counts of each YouTubers. Then, linear regression and neural networks are trained to classify and find the correlation between these variables. Finally, Matplotlib, google chart, and Tableau are utilized to visualize the result based on video categories and geographical locations. The accuracies of linear regression and neural network models are verified through the R-squared estimation. Both linear regression and neural network models show the trending types of videos and a positive correlation between the number of viewers and subscribers. The experimental results show a remarkable user's tendency of watching films and listening to music, a concentration of YouTube users from India and the U.S., and propose targeted Covid-19 prevention propaganda based on the above two characteristics.**

Keywords-YouTube; Deep Learning; Linear Regression; Neural Network; Machine Learning

I. INTRODUCTION

YouTube data analysis intends to study the video preferences of YouTube users under the current COVID society, based on the view counts and the number of subscriptions of specific YouTube channels. It achieves a systematic video classification to identify the most representative categories for data analysis. The different countries of YouTube videos are also considered a crucial element to accomplish the geographical location analysis of video audiences. Combining with the YouTube video classification and particular machine learning algorithms, such geographical analysis successfully produces the distribution of YouTube video preference and informative and pertinent measures of spreading COVID prevention.

More effective and directed propagating content can be formulated for COVID prevention by analyzing YouTube data. Acknowledging the video tendency of YouTube users in different locations, people can integrate specific information about the pandemic in each country into promoting COVID prevention within videos from distinct regions. In other words, this video tendency enables viewers to learn local COVID precautions while watching their favorite YouTube videos. In addition to COVID prevention, YouTube data analysis outcomes also promote positive information and urgent news to target populations. Furthermore, based on the statistical analysis of YouTube video categories, YouTubers can find out the most welcomed video at the present stage, thus improving the popularity and heat of their videos and increasing their number of subscriptions by following the trend. The YouTube data analysis aims to find the correlation between view counts and subscriptions for different YouTube channels.

In this paper, we use Linear Regression and Neural Network to analyze the YouTube data for YouTube video preference analysis. First, the YouTube data are extracted from the YouTube API console and filtered through the pandas' package[1]. Then the data are plugged into machine learning models, including both linear regression and neural network. The results of models are analyzed visually using data visualization tools Matplotlib and tableau as well as statistically by R-squared and the best fit lines. The experimental results present a positive correlation between the number of subscriptions and views of YouTube Channel, a high concentration of YouTube video views and subscriptions in India and the U.S, and a video preference of watching films and listening to music.

II.METHOD

This section describes our method for YouTube data analysis for video preferences based on labeled data. Figure 1 shows the whole architecture of our method. As shown in figure 1, we first process the YouTube data (Sec. A). Then, the processed data are used for both linear regression and neural network model construction (Sec. B), followed by the statistical results and result estimation (Sec. C). Finally, the data visualization is completed for an intuitive exhibition of YouTube data preferences (Sec. D).

978-1-6654-8476-3/22/$31.00 ©2022 IEEE
DOI 10.1109/BDICN55575.2022.00055

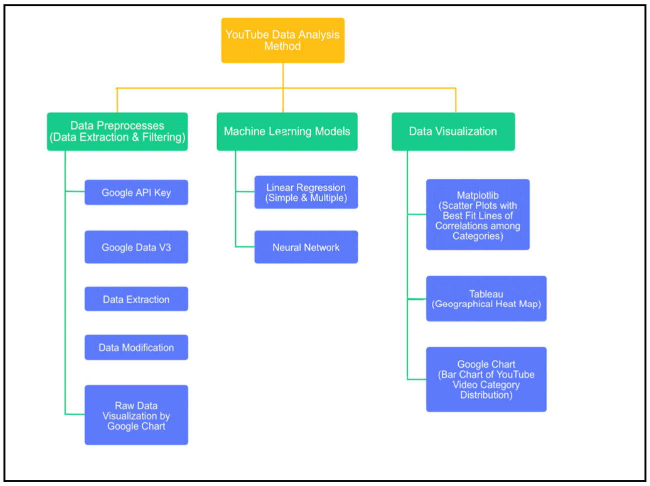
YouTube Data Analysis Method
Data Preprocesses (Data Extraction & Filtering)
Machine Learning Models
Data Visualization
Google API Key
Google Data V3
Data Extraction
Data Modification
Raw Data Visualization by Google Chart
Linear Regression (Simple & Multiple)
Neural Network
Matplotlib (Scatter Plots with Best Fit Lines of Correlations among Categories)
Tableau (Geographical Heat Map)
Google Chart (Bar Chart of YouTube Video Category Distribution)

Figure 1. Flow Chart of YouTube Data Analysis.

A.Data Pre-processing

YouTube data extraction is achieved through accessing YouTube data API, a platform containing statistical data of YouTube channels[4]. The API key is required as a certification to access such data, which can be retrieved from Google Developers Console. After the YouTube data API v3 is invoked, variables like channel ID, number of videos, views, subscribers, channel countries, and video categories are selected into the data frame. Such a data frame enables the data analysis by linear regression and neural network and the study of the characteristics of each video genre. Since the YouTube API only provides the lists of the combination of categories and subcategories of each channel's videos, the data filtering is processed to sort out the major categories from the lists based on keywords of each main category.

The raw data is classified into distinct categories, which is further enumerated within each category and visualized through Google Chart[2]. Figure 2 presents the distribution of all the categories of channels, with music having the highest ratio of videos(28.1%) and entertainment having the second-largest ratio(26%).

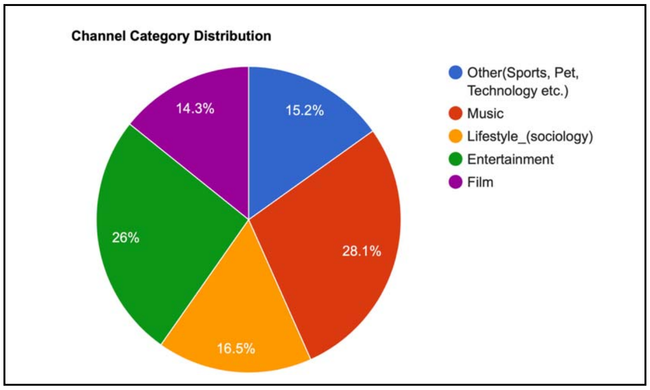

Figure 2. Distribution of Channel Category.

B.Model Construction

This subsection covers our model construction, including linear regression and neural networks. Both models are utilized for YouTube analysis because the high performance of the neural network accompanies risks due to its unsupervised property[9]. The unsupervised learning trait of the neural network adds a "black box" property to this model, creating difficulties for researchers in some particular circumstances. Therefore, the supervised linear regression is also utilized in the YouTube study as a traditional machine learning algorithm to reconcile such challenges since it is less risky, and its inputs are controllable by researchers compared to the neural network. Using linear regression only produces linear correlation among variables and is more sensitive to outliers. In other words, linear regression performs worse when dealing with unlabeled data than the neural network's performance in such an area. Therefore, the combination of the two algorithms, without losing performance and generality, allows visualizing the user's video preferences based on geographical locations and different categories more conclusively.

1）Linear Regression

Linear regression is used to fit the independent variable, the number of subscribers, the dependent variable, and the number of views within each video category. Linear regression algorithm fits the data set with coefficients to minimize the residual sum of squares between each point on the graph[6]. Such linear correlation between the number of subscribers and viewers is visualized via the scatter plot in Matplotlib as figure 3. Similarly, multiple linear regression is also utilized to obtain the association between independent variables, the number of subscriptions, videos, and the number of views dependent variable[7]. Note that since both independent variables and dependent variables contain large numbers, the training and predicting of machine learning models take the logged value of each variable, which does not affect the correlation between independent and dependent variables as they shrink to the same degree.

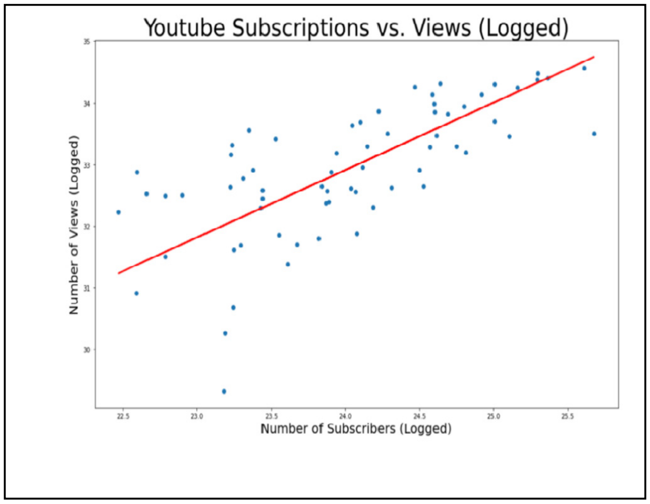

Figure 3. Linear Regression Result of YouTube Subscriptions vs. Views.

2）Neural Network

Perceptron classifier is used in neural networks to predict the number of views by two independent variables, subscriptions and videos[3]. The MLP Regressor is invoked for views' prediction within each genre and trains the model in steps of the partial derivatives of the loss function to update the parameters. Random number generation for weights and bias initialization and train-test split is used to randomly assign data to test and training groups for the following training of the neural network algorithm. In Multi-layer Perceptron, the regularization term is added to the loss function to shrink the model parameters for preventing overfitting. Hyperparameters such as the maximum number of iterations and learning rates are altered to optimize the result. The testing set is then being compared with the actual class to find out how the network performs. Since there are multiple types of videos in the data set, the neural network performs well when dealing with nonlinearities.

C. Visualization

Different categories of YouTube videos are analyzed by visualizing the data as well as statistics. The intercepts and slopes of best fit lines represent the sizes of the population of the initial audience and the abilities to attract the new audience of different categories.

1)Matplotlib

Matplotlib is applied in YouTube data analysis to visualize the data assessment result by drawing the line chart and bar chart for actual and model-predicted views' value comparisons. Moreover, the combination of scatter plots and lines provides intuitive results of the trends of YouTube data when seeking the correlation between the number of subscriptions and views within and among categories.

2)Tableau

Tableau is a data visualization tool that provides an accessible way to see trends and patterns in data, which is applied to visualize the distribution and density of the number of views and subscribers in terms of geographical locations for finding the most frequent YouTube users in a worldwide scale [8]. Tableau is utilized to analyze the geographical densities of the uploaded video by YouTubers, the views of videos of each channel, as well as the number of subscribers of each channel. Such geographical heat maps of YouTube data are generated by plugging in YouTube variables, including the number of views, subscriptions, videos, and countries.

3)Google Chart

Google chart is used to offer better visualization of the distributions of each category, which can be helpful in terms of analyzing the data. Google chart performs well in integrating data from other Google products, including the coding platform Google Collab, which is used in the category filtering section of YouTube data analysis. In the YouTube data analysis, the Google chart is employed for indicating the proportions of the number of videos of each YouTube video category by pie charts.

III.RESULTS AND DISCUSSION

This section introduces the experimental settings (Sec. 3.1), the evaluation results of the linear regression and the neural network models, and the analysis of YouTube video preferences and characteristics based on categories and geographical locations (Sec. 3.2).

A.Experimental Settings

Dataset. The experimental settings of YouTube data analysis include the meaning of YouTube dataset's variables, which act as input for machine learning models. Table 1 shows the meaning and the variables of YouTube statistics.

Evaluation Metric. A similar R-squared is calculated for the model assessment of multiple linear regression and the neural network.

TABLE I. YOUTUBE DATASET

Variables	Meaning
Subscription	The number of subscriptions of each channel.
View	The number of views of each video.
Channel ID	The ID number of specific channel (Used for extracting statistics from YouTube API console)
Country	The country of each channel.
Category	The category of each video.
Video	The number of videos a channel uploaded.

B.Results and Analysis

Table 2 shows the comparison of the accuracy of simple linear regression, multiple linear regression, and neural networks. For simple linear regression, the r-squared of the predicted values of the dependent variable (the number of viewers) are calculated. Since the R squared is shown with a value of about 0.47, it is confident to conclude that the trained model explains about 47 percent of the variation in the YouTube numbers of views. Figure 4 also shows that the larger the sample size is, the weaker the capability of each model in explaining the variation of YouTube data. Yet, the 45 percent of variance capturing ability in large sample size is enough for concluding a relatively accurate predicting outcome.

TABLE II. R-SQAURED TABLE OF LINEAR REGRESSION AND NEURAL NETWOK

	Simple Linear Regression	Multiple Linear Regression	Neural Network
123 channels	84.7%	94.8%	87.5%
489 channels	46.9%	42.1%	45.9%

The best fit lines generated from simple linear regression show that subscription and views are positively correlated. Among the best fit lines of different categories shown in figure 4, the film genre possesses the highest values of coefficients, meaning that when subscribers increase at the same rate, the viewer of the film increases with the highest rate compared to other major categories. Moreover, the result also shows that music obtains the highest intercept, which implies that the music channel has the largest audience population at its set-up phase.

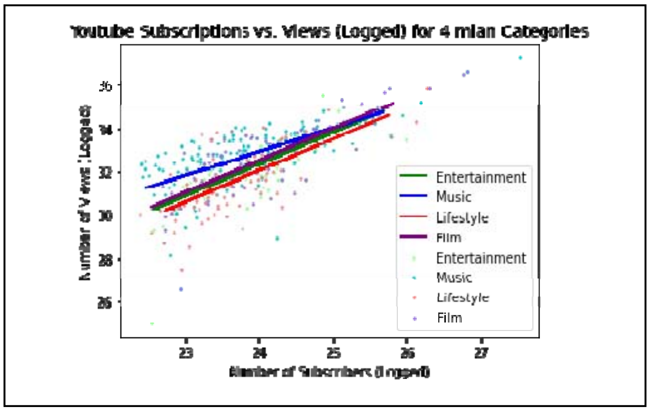

Figure 4. Linear Regression Results of Multiple YouTube Video Categories.

In the visualization from Tableau, the darker the color, the more videos, subscribers, and views there are[8]. Based on the heat map, it has been found that the United States and India are the most active countries in posting videos on YouTube and have the largest number of subscribers and views[10]. Figure 5 shows that the United States has the highest number of videos, showing that there are 1,327,720,000 subscribers, 36,875 videos, and 718,596,611,074 views.

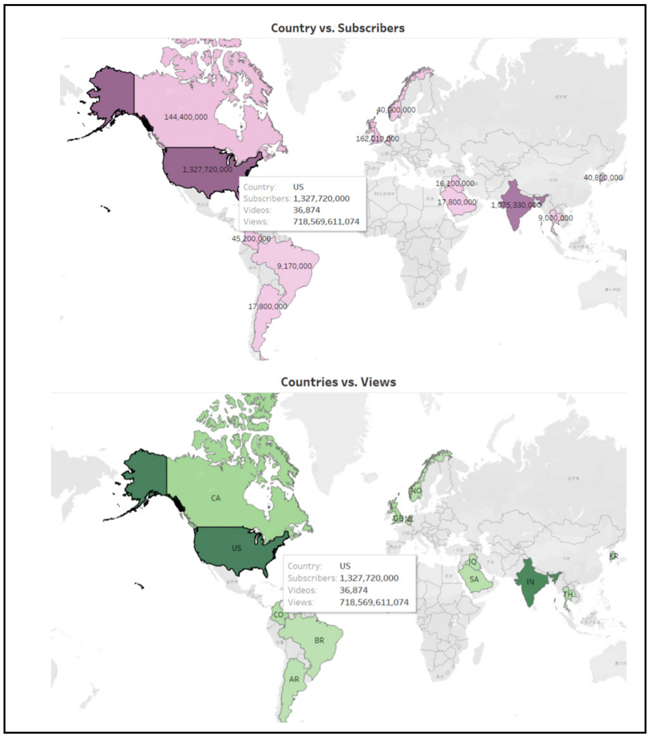

Figure 5. Geographical Distribution of YouTube Videos Subscribers and Views.

IV.CONCLUSION

This paper analyzes YouTube data through linear regression and the neural network. The analysis keeps pace with the current social phenomena, extracts the latest YouTube data to trace the adjusted correlation between people and social media under the influence of Covid-19. The research focuses on the YouTube user's tendency of different types of videos as well as the geographical distribution of YouTube users and video watching. The numerical experimental results verify the effectiveness of the proposed methods described above. Our analysis of YouTube data illustrates a remarkable user's tendency of watching films and listening to music, and a concentration of YouTube users from India and the U.S., offering a detailed measure by inserting Covid-19 prevention propaganda within films, music, as well as focusing on the local prevention of India and the United States.

In the future, the study will focus on the geographical preferences of YouTube videos and the influence of comments on the heat of YouTube videos through the Natural Language Toolkit. The study of geographical video preference and video heat aims to achieve artificial assessment of video potential, which is able to benefit YouTubers by giving proper advice on the most profitable and heated videos based on the preferences of their audiences.

Refereces

[1]Saini, V. (2020, December 19). Extracting YouTube data with python using API. Present Slide.

[2]Google. Visualization: Pie chart | charts | google developers. Google.

[3]Sklearn.neural_network.MLPClassifier.scikit. https://scikitlearn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html.

[4]Babikov, I. (2018, November 15). YouTube channels ~100000. Kaggle. https://www.kaggle.com/babikov/youtube-channels-100000.

[5]Mapping in Tableau. Tableau. https://help.tableau.com/current/pro/desktop/en-us/maps_build.htm.

[6]Lamsal, S. (2021, February 22). Multiple linear regression: Sklearn and Statsmodels. Medium. https://codeburst.io/multiple-linear-regression-sklearn-and-statsmodels-798750747755.

[7]Linear Regression (python implementation). GeeksforGeeks. (2021, September 14). https://www.geeksforgeeks.org/linear-regression-python-implementation/.

[8]Maps. Tableau. https://www.tableau.com/solutions/maps.

[9]1.17. neural network models (supervised). Scikit. https://scikitlearn.org/stable/modules/neural_networks_supervised.html.

[10]Tableau Map Layers (2021, February 28). datavis.blog. https://datavis.blog/2020/12/02/tableau-map-layers/.