Bulletin of Electrical Engineering and Informatics
Vol. 8, No. 4, December 2019, pp. 1502~1506
ISSN: 2302-9285, DOI: 10.11591/eei.v8i4.1630

A regression approach for prediction of Youtube views

Lau Tian Rui[1], Zehan Afizah Afif[2], R. D. Rohmat Saedudin[3], Aida Mustapha[4], Nazim Razali[5]
[1,2,4,5]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
[3]School of Industrial Engineering, Telkom University, 40257 Bandung, West Java, Indonesia

Article Info

Article history:

Received Apr 30, 2019
Revised Jun 18, 2019
Accepted Jul 10, 2019

Keywords:

Prediction
Regression
Social media
Social networking
Youtube views

ABSTRACT

YouTube has grown to be the number one video streaming platform on Internet and home to millions of content creator around the globe. Predicting the potential amount of YouTube views has proven to be extremely important for helping content creator to understand what type of videos the audience prefers to watch. In this paper, we will be introducing two types of regression models for predicting the total number of views a YouTube video can get based on the statistic that are available to our disposal. The dataset we will be using are released by YouTube to the public. The accuracy of both models are then compared by evaluating the mean absolute error and relative absolute error taken from the result of our experiment. The results showed that Ordinary Least Square method is more capable as compared to the Online Gradient Descent Method in providing a more accurate output because the algorithm allows us to find a gradient that is close as possible to the dependent variables despite having an only above average prediction.

Corresponding Author:

Zehan Afizah Afif,
Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia,
86400 Parit Raja, Batu Pahat, Johor, Malaysia.
Email: afizah@uthm.edu.my

1. INTRODUCTION

The advancement of technology results in the growing amount of online entertainment website. As it becomes apparent that entertainment is shifting toward new media, people loves to visit a website which bring them fun contents or useful knowledge during their leisure time, and YouTube is always be their best choice. In fact, since its launching in 2005, YouTube has become one of the world's most powerful digital media platforms [1]. YouTube, as a video sharing website, provide a platform for their users to enjoy watching videos or also create their own videos to entertain audiences, for education use, advertisement and others. Due to the needs of the entertainer in Web 2.0 application, the popularity of web content has soon becoming the hot topic around the world, furthermore it also helps to generate income for a YouTuber.

Research using YouTube data has been also receving growing attention such as comments annotation [2], opinion mining [3], sentiment analysis [4] or social media analytics [5] for YouTube videos. There are several articles and journals that have previously been published in relation to investigating the popularity of YouTube videos. In [6], different models were presented for determining YouTube view-count, which showed the viral trends and potential population growth. Based on these models, they proposed an automatic classification of the YouTube videos that classify a video into one of four categories, which were viral and fixed population; viral and growing population; non-viral fixed population; and non-viral growing population. View counts are extremely important in modeling and characterizing viewers of YouTube videos [7] as well as in analyzing the popularity patterns of user shared content [8, 9].

The work by [10] concluded that while some YouTube videos become viral and instant hits, majority of the videos are only experiencing limited interest. Two methods were proposed to predict video popularity based on the content as input variable abd then the daily samples of the content popularity are measured up to a given reference date. Other than the content, variable list includes the daily samples of number of comments, number of ratings and number of users who "favourited" the video as these information are readily available in the video statistics panel. However, since the variables are found to be highly correlated with the number of views, the variables do not affect the prediction outcome from a linear regression.

HIPie (Hawks Intensity Process Insights Explorer), a software tool developed by [11], had the ability to analyse and predict the future popularity of YouTube videos. HIPie was developed based on the Hawkes Intensity Process (HIP) [12], which is an interactive web-based that allows users to reason about the popularity and the virality of Youtube videos. Kong et al. stated that the software is able to predict the popularity of YouTube videos through the analysis of several factors. For any video, HIPie depicts several popularity series: observed, fitted and forecasted by HIP. HIPie enables users to comparatively analyse videos using the endo-exo map, by showing the view count and the number of shares they receive, alongside the exogenous sensitivity and the endogenous reaction. The software can also identify videos that have the potential of going viral, but are yet to.

Nonetheless, the recent viral YouTube video seems to have varied. There is no specific rule-of-thumb for a viral video, it basically describes a phenomenon in which a video clip become highly popular through rapid, user-sharing content via the Internet, which in this case a YouTube platform. This type of marketing distribution is led by users through what is known as a participatory popular culture online [13]. To date, the literature also showed different research on predicting Youtube views based on popular domain such as educational videos [14, 15], video games live streaming [16, 17], and politics [18] among few.

In this paper, we explore the factor that influence YouTube video's view counts and attempt to estimating it using regression method. The remainder of this paper proceeds as follows. Section 2 presents the research methodology, Section 3 presents the results, and finally Section 4 concludes with some plan for future work.

2. RESEARCH METHOD

In this experiment, our main objective is to predict the potential total view count of a YouTube video as accurate as possible based on several influential factor. To do so, we have decided to apply one of the predictive modeling techniques, which is the regression technique in our experiment. We will be using regression technique to model the mathematical correlation between our independent variables, which are the attributes in the dataset we have acquired, and the dependent variable, in this case it is the amount of viewership of a YouTube video. After figuring out the pattern and the relationship between the said variables, we are then able to predict the future value of the dependent variable.

One of the key benefits that regression analysis offers is that it indicates the strength of impact of multiple independent variables on a dependent variable. This will allow us to compare the outcome when a variable has its value changed. For example, the result of this experiment will show how the channel subscriber count will affect the total view count of a YouTube video.

2.1. Dataset

Dataset that have chosen for this project is the trending YouTube Video Statistics (https://www.kaggle.com/datasnaek/youtube-new) from daily statistics for trending YouTube videos in Kaggle [19]. Kaggle is a website that provides dataset for data scientists and machine learners. It allows us to download data sources in both CSV and JSON format. This dataset includes several months of data on daily trending YouTube videos from regions such as the USA, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan and India. Data from each region is stored in a separate file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count. The excerpt of the dataset is shown in Table 1.

Table 1. Trending statistics of YouTube videos

Likes	Dislikes	Comment Count	Comment Disabled	Rating Disabled	Video Error / Removed	Views
1,094,557	8,876	65,275	False	False	False	16,146,848
252,432	10,936	9,291	False	False	False	16,154,588
211,395	16,998	225,681	False	False	False	16,165,607
170,269	5,343	12,663	False	False	False	16,178,195
405,952	10,025	28,779	False	False	False	16,187,620

Figure 1 and Figure 2 shows data visualization based on the Ordinary Least Square Method and Online Gradient Descent, respectively. As seen in Figure 1 and Figure 2, the visualizations illustrate the number of instances (rows) in the datasets as well as the number of variables (columns). Basically it visualized the data in terms of their statistical properties such as mean, median, min, max, standard deviation, unique values, missing values as well as type of features of variables.

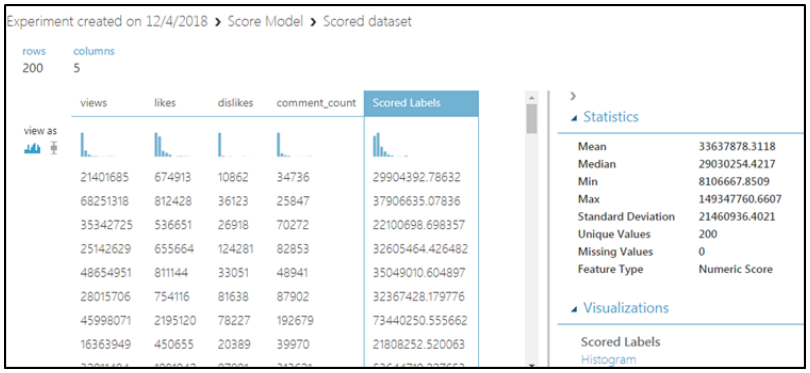

Figure 1. Data visualization using ordinary least square method

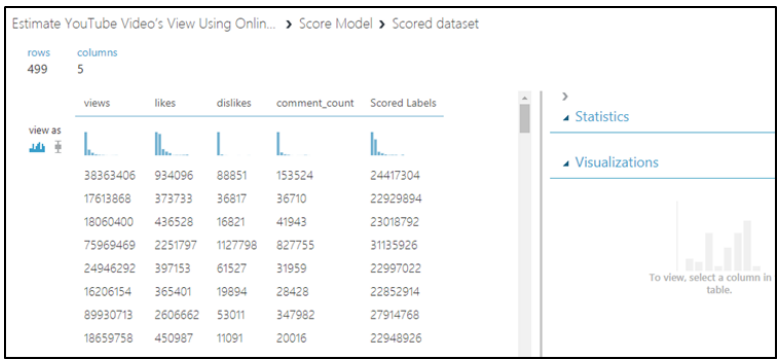

Figure 2. Data visualization using online gradient descent

2.2. Algorithms

The evaluation metrics used in the experiments are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Coefficient of Determination. Mean Absolute Error (MAE) calculates the average difference between the estimated value and the predicted value. MAE is shown in (1).

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|x_i - y_i| \quad (1)$$

Root Mean Square Error (RMSE) is the squared root of the average difference between the estimated value and the predicted value. RMSE is shown in (2).

$$MAE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}|x_i - y_i|} \quad (2)$$

Coefficient of Determination is the number that determines whether a statistical model fits a data set. Coefficient of Determination is shown in (3).

$$R^2 = 1 - \frac{SSE_{reg\_line}}{SSE_{mean\_y}} \quad (3)$$

2.3. Algorithms

For this project, we will be using two algorithms which are the Ordinary Least Squares regression (OLS) [20] and the Online Gradient Descent Algorithms [21]. Ordinary Least Square regression, also commonly known as just linear regression, is used to minimize the sum of square of differences between the dependent and predicted value. It is used to predict the output's values for new samples. The online gradient descent, also known as Stochastic Gradient Descent [22] is used to get a stochastic approximation of a gradient descent optimization. Stochastic means that it selects randomly instead of in order in a training set.

For our project, there will be four variables. The independent variables will be the like count, dislike count and the comment count of YouTube videos while the dependent variable will be the view count. Using the OLS algorithm allows us to reduce positive and negative residual cancelling each other and finding a gradient that is close as possible to the dependent variables. The OLS algorithm allows us to illustrate our predictions on average and create a conclusion that follows from the regression line passing through the sample means.

Using the Stochastic Gradient Descent algorithm, while it tends to be noisier due to the usage of only one selection from a training set, it still gives us the minimum and it has a much shorter training time. However, due to it's stochastic nature, the path towards the global minimum is not as direct as a regular Gradient Descent algorithm would be, the gradient may instead possess a more 'zig-zagged' pattern along the gradient. The Stochastic Gradient Descent algorithm is designed for use in large sample sizes which is ideal for our project as our project concerns the view count of YouTube videos as the dependent variable. As the view count will normally be a large value, the algorithm will prove efficient in generating an output in a short amount of time.

3. RESULTS AND DISCUSSION

The purpose of the experiments is to compare the performance of Ordinary Least Squared Method and Online Gradient Descent Method in predicting the views of a YouTube Video. After running the experiment, the evaluation results are shown in Table 2.

Table 2. Summary of results

Metrics	Ordinary Least Squared Method	Online Gradient Descent Method
MAE	10212065.610017	14219674.511022
RMSE	14567739.474194	26761202.017444
R2	0.681229	-0.033356

The results showed that the MAE and RMSE using Ordinary Least Square Method are 10,212,065.610 and 14,567,739.474 respectively. The MAE and RMSE when using Online Gradient Descent Method are 14,219,674.511 and 26,761,202.017 respectively. Just looking at the numbers indicates that the MAE and RMSE of both methods are absurdly high. However, by comparing both experiments, Ordinary Least Square Method has lower MAE and RMSE. In other words, Ordinary Least Square Method has a higher accuracy when it comes to predicting YouTube video views as compared to Online Gradient Descent Method because the difference between the estimated and the real values of Ordinary Least Square Method is lesser.

Next, the R-Squared obtained when using Ordinary Least Square Method is 0.681, which means 68.1% of the total variation is explained by the regression line using the independent variable. Meanwhile, the R-Squared from using Online Gradient Descent Method is -0.033. A negative R-Squared is rare and in this case, it shows that this method is not viable for this experiment and it does not fit the data at all.

4. CONCLUSION

Estimating YouTube's view using Regression Method is possible using Ordinary Least Square Method compared to Online Gradient Descent Method but the prediction is only above average. This may be caused by the attribute selected or we may use data integration method in preprocessing phase to increase the accuracy of the prediction outcome. As stated in the algorithm section, despite being able to generate the output in a short amount of time, the generated output tends to be noisier, providing a less accurate result. This shows that the Ordinary Least Square method is more capable of providing a more accurate output due to the algorithm allowing us to find a gradient that is close as possible to the dependent variables despite

having an only above average prediction. In the future, we hope to inevestigate the impact of using multiple regression models such as in [23] and [24] or other regression-based models [25].

ACKNOWLEDGEMENTS

This paper is funded by International Grant Scheme vot W004 at Universiti Tun Hussein Onn Malaysia.

REFERENCES

[1] Burgess J, Green J. YouTube: Online video and participatory culture. John Wiley and Sons. 2018 Aug 28.
[2] Li T, Lin L, Choi M, Fu K, Gong S, Wang J. "Youtube AV 50K: an annotated corpus for comments in autonomous vehicles". arXiv preprint arXiv:1807.11227. 2018 Jul 30.
[3] Rinaldi E, Musdholifah A. "FVEC-SVM for opinion mining on Indonesian comments of youtube video," *2017 International Conference on Data and Software Engineering (ICoDSE)*, Palembang, 2017, pp. 1-5.
[4] Asghar MZ, Ahmad S, Marwat A, Kundi FM. "Sentiment analysis on Youtube: A brief survey". arXiv preprint arXiv:1511.09142. 2015 Nov 30.
[5] Thelwall M. "Social media analytics for YouTube comments: potential and limitations". *International Journal of Social Research Methodology*. 2018 May 4;21(3):303-16.
[6] Richier C, Altman E, Elazouzi R, Altman T, Linares G, Portilla Y. "Modelling view-count dynamics in youtube". arXiv preprint arXiv:1404.2570. 2014 Apr 9.
[7] Aggrawal N, Arora A, Anand A. "Modeling and characterizing viewers of You Tube videos". *International Journal of System Assurance Engineering and Management.* 2018, Apr 1:1-8.
[8] Tanaka T, Ata S, Murata M. "Analysis of Popularity Pattern of User Generated Contents and Its Application to Content-Aware Networking," *2016 IEEE Globecom Workshops (GC Wkshps)*, Washington, DC, 2016, pp. 1-6.
[9] Murata M, Kitade Y. "Analyzing Popularity Dynamics of YouTube Content and its Application to Content Cache Design". 2015.
[10] Pinto H, Almeida JM, Gonçalves MA. "Using early view patterns to predict the popularity of youtube videos". *In Proceedings of the sixth ACM international conference on Web search and data mining* 2013 Feb 4 (pp. 365-374).
[11] Kong Q, Rizoiu MA, Wu S, Xie L. "Will This Video Go Viral? Explaining and Predicting the Popularity of Youtube Videos". arXiv preprint arXiv:1801.04117. 2018 Jan 12.
[12] Rizoiu MA, Xie L, Sanner S, Cebrian M, Yu H, Van Hentenryck P. "Expecting to be HIP: Hawkes intensity processes for social media popularity". *In Proceedings of the 26th International Conference on World Wide Web.* 2017, pp. 735-744.
[13] Burgess J. "All your chocolate rain are belong to us? Viral Video, YouTube and the dynamics of participatory culture". In Video vortex reader: Responses to YouTube. 2008, pp. 101-109. Institute of Network Cultures.
[14] Shoufan A, Mohamed F. "On the likes and dislikes of youtube's educational videos: A quantitative study". *In Proceedings of the 18th Annual Conference on Information Technology Education* 2017 Sep 27 (pp. 127-132).
[15] Shoufan A. "Estimating the cognitive value of YouTube's educational videos: A learning analytics approach". *Computers in Human Behavior*. 2019 Mar 1; 92:450-8.
[16] Kaytoue M, Silva A, Cerf L, Meira Jr W, Raïssi C. "Watch me playing, I am a professional: A first study on video game live streaming". *In Proceedings of the 21st international conference on World Wide Web ACM*. 2012 Apr 16 (pp. 1181-1188).
[17] Lessel P, Mauderer M, Wolff C, Krüger A. "Let's Play My Way: Investigating Audience Influence in User-Generated Gaming Live-Streams". *In Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video* 2017 Jun 14 (pp. 51-63).
[18] Church SH. "YouTube politics: YouChoose and leadership rhetoric during the 2008 election". *Journal of Information Technology & Politics*. 2010 May 18; 7(2-3):124-42.
[19] Mitchell J. Trending Youtube Video Statistics and Comments. Kaggle, Kaggle Inc. 2017 Aug.
[20] Hutcheson GD. Ordinary least-squares regression. L. Moutinho and GD Hutcheson, The SAGE dictionary of quantitative management research. 2011:224-8.
[21] Hazan E, Rakhlin A, Bartlett PL. "Adaptive online gradient descent". *Advances in Neural Information Processing Systems* 2008 (pp. 65-72).
[22] Bottou L. "Stochastic gradient descent tricks. Neural networks: Tricks of the trade". Springer, Berlin, Heidelberg. 2012 (pp. 421-436).
[23] Suyono H, Hasanah RN, Setyawan RA, Mudjirahardjo P, Wijoyo A, Musirin I. "Comparison of Solar Radiation Intensity Forecasting Using ANFIS and Multiple Linear Regression Methods". *Bulletin of Electrical Engineering and Informatics*. 2018,7(2):191-8.
[24] Suparta W, Putro WS. "Comparison of tropical thunderstorm estimation between multiple linear regression, Dvorak, and ANFIS". *Bulletin of Electrical Engineering and Informatics*. 2017, 6(2):149-58.
[25] Darlington RB. "Regression and linear models". New York: McGraw-Hill; 1990 Dec.