Final Report

Predictive Analytics (ISQS 6349)

Marcin Grzechowiak, Alex Cathey, Mychael Solis-Wheeler, Mikaela Pisani, Roger Valdez

INTRODUCTION

The main idea of this project is to analyze the business model of TED Talks. Our audience is video publishing companies, company sponsors, and TED Talk presenters interested in maximizing their views. In order to do that, we analyze factors that can affect the number of unique *views* per video. This is a valuable effort because in the world of YouTube, more *views* means more profit for the publisher and more engagement for the presenters. We are interested in analyzing what can influence the decisions that the presenters and companies makes in relation to the date that the video is published. The analysis is based on a dataset from kaggle.com, which contains information about the amount of unique *views* per video, the duration of the video, the topics that it covers, the languages that it is translated to, etc.

We decided to focus our attention in the following questions:

1. **PROBLEM 1**: Relationship between number of unique *views* and *duration*.
2. **PROBLEM 2**: Relationship between type of tags and number of *views.*
3. **PROBLEM 3**: Relationship between type of tags and publish date, specified as the months between May and August.

Observations

- Number of *views* is a measure that is unique per individual user
- *Duration* is measured in seconds
- Type of *tags* means the type of topic of the talk
- The date that the video is published, *published_date* is transformed into a binary variable that determines whether the date corresponds to a vacation period or not (May-August)

PROBLEM 1

a. Question: What is the relationship between number of unique *views* and the *duration* of a video?
b. Background:
    i. Previous researchers have observed that the length of a Youtube video is negatively related to unique *views* per YouTube video (Park et al., 2016). Therefore, we hypothesize that as video *duration* increases, unique *view* count decreases. Additionally, previous researchers have observed that watchtime (measured in *views*) percentage is impacted by video duration (Gueo et al., 2014). Thus, this background evidence further supports our hypothesis.
    ii. We are interested in which measure of video performance most closely dictates a videos' success, being determined as a higher number of views on a published

video. It is important to note the correlation between the variable of language translations and views. Generally, we can say that the more languages a video is translated into, the wider the audience for video viewership globally. A broad audience yields more opportunity for sharing the video via word of mouth or through social media (Boppolige and Gurtoo, 2017).

c. Empirical Results & Discussion

We run the following models to test the influence of *duration* for the dependent variable *views*:

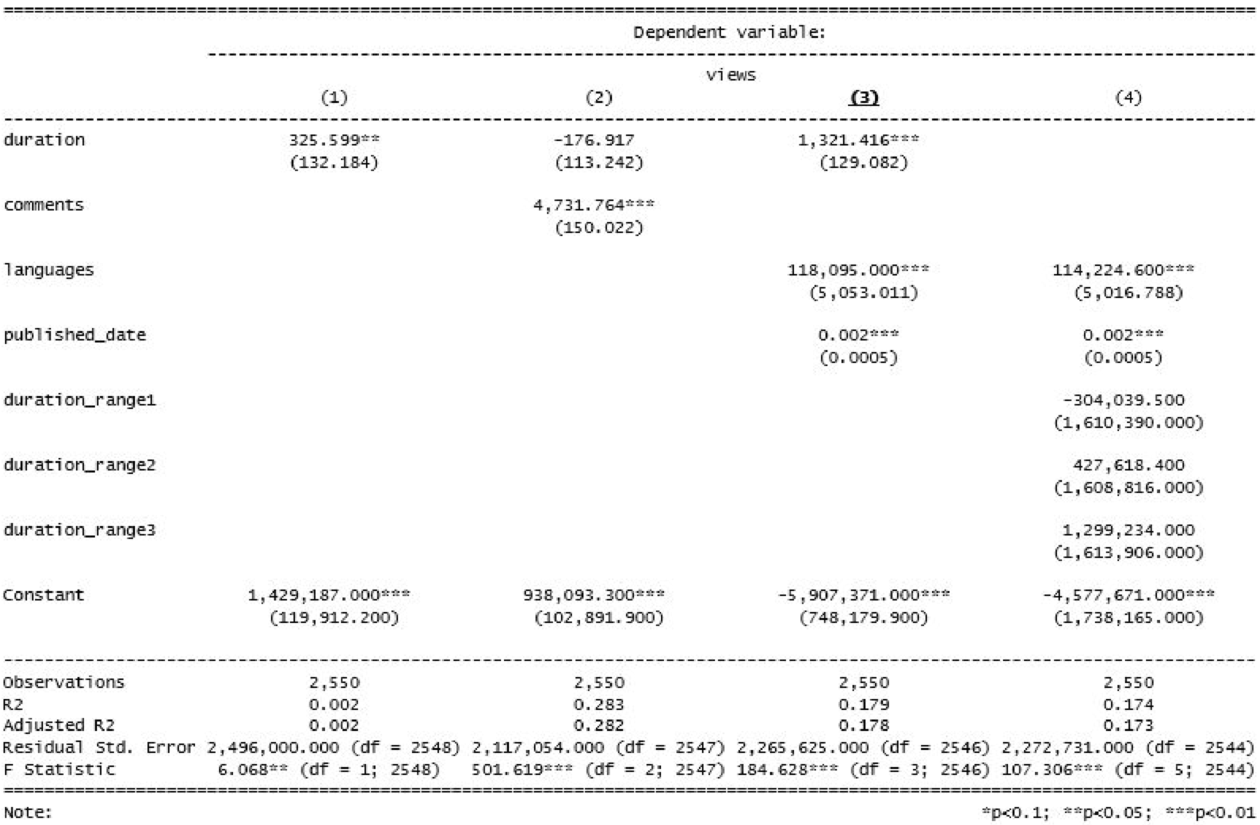
i. Model 1: Linear Regression Model including only *duration* as the independent variable.

ii. Model 2: Multiple Linear Regression Model: including *duration* and *comments* as independent variables.

iii. Model 3: Multiple Linear Regression Model: including *duration*, *languages* and *published_date* as independent variables.

iv. Model 4: Multiple Linear Regression Model: including *duration*, *languages, published_date,* and *duration_range[1-3]* as independent variables. (*duration_range* variables are binary)

The results obtained are presented below:



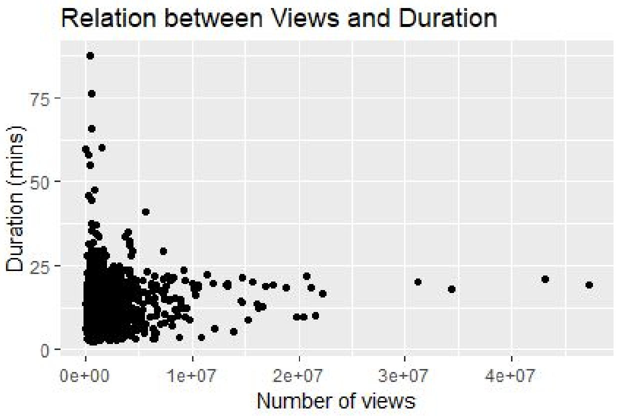
```
==============================================================================================================
                                                  Dependent variable:
                     -----------------------------------------------------------------------------------------
                                                         views
                             (1)                    (2)                     (3)                     (4)
--------------------------------------------------------------------------------------------------------------
duration                 325.599**               -176.917              1,321.416***
                        (132.184)               (113.242)               (129.082)

comments                                       4,731.764***
                                                (150.022)

languages                                                            118,095.000***          114,224.600***
                                                                       (5,053.011)             (5,016.788)

published_date                                                          0.002***                0.002***
                                                                        (0.0005)                (0.0005)

duration_range1                                                                              -304,039.500
                                                                                            (1,610,390.000)

duration_range2                                                                               427,618.400
                                                                                            (1,608,816.000)

duration_range3                                                                              1,299,234.000
                                                                                            (1,613,906.000)

Constant              1,429,187.000***        938,093.300***         -5,907,371.000***       -4,577,671.000***
                        (119,912.200)          (102,891.900)           (748,179.900)          (1,738,165.000)

--------------------------------------------------------------------------------------------------------------
Observations               2,550                  2,550                   2,550                   2,550
R2                         0.002                  0.283                   0.179                   0.174
Adjusted R2                0.002                  0.282                   0.178                   0.173
Residual Std. Error 2,496,000.000 (df = 2548) 2,117,054.000 (df = 2547) 2,265,625.000 (df = 2546) 2,272,731.000 (df = 2544)
F Statistic          6.068** (df = 1; 2548)   501.619*** (df = 2; 2547) 184.628*** (df = 3; 2546) 107.306*** (df = 5; 2544)
==============================================================================================================
Note:                                                                          *p<0.1; **p<0.05; ***p<0.01
```

Variables included in the above regression models:

1. *Duration*: measured in seconds, shows the length of a published video
2. *Views:* measured in the amount of unique views per individual user
3. *Comments*: amount of comments that the video has
4. *Languages*: the count of translated languages on one video
5. *Published_date*: the date the video was posted to public viewing
6. *Duration_range*: is a binary variable that determines if the duration of the video is in one of the three ranges (<10 minutes, 10-20 minutes or >20 minutes)

Results & Interpretation

1. Model 1: We ran a linear regression model because we are interested in the direct influence that *duration* has on *views*. Prior to running the first regression, we expected to see a negative correlation between *duration* of video and the number of unique *views*. After running the regression, we determined that there were explanatory variables in the dataset that if excluded would result in an omitted variable bias.
2. Model 2: We then included the independent variable *comments* and ran a consecutive regression analysis. After running the new model, we discussed the possibility that there is a simultaneous causality bias between the variables *comments* and *views*, because we cannot determine if *comments* affects *views* or views affects comments. Thus, we decide to remove the variable *comments* from the model.
3. Model 3: Adding the variables *languages* and *published_date,* the regression analysis results indicate the presence of an omitted variable bias in the first model. We can see this as the coefficient changes significantly as we add new variables. The regression results contradict our original hypothesis that *views* would decrease as *duration* increases. The results show that there is a positive correlation between *duration* and *views.*
4. Model 4: In order to analyze the effect of duration on views, we decide to run a regression using 3 binary variables for different ranges between duration. The coefficients obtained are not significant for the ranges of duration.
5. Analyzing the graph presented below, we can see that the positive correlation between *duration* and *views* is only present up until a certain point, where there is a constant return to scale. Since after this point the number of views continues increasing, we assume that there are other variables that affect *views*, such us *languages* and *published_date*.



6. Chosen model _(3): We decide that between these models the one that represents better the reality is the one that includes the independent variables *duration*, *languages* and *published_date*, which are significant and positively correlated with the amount of *views*.
   $views = \beta_0 + \beta_1 * duration + \beta_2 * languages + \beta_3 * published\_date$
   Getting 1321 for the coefficient for *duration*, which means that if duration increases 1 second, *views* increases by 1321. This positive correlation between *duration* and *views* is up until a certain point where other variables are influencing the number of *views* per video, but the *duration* does not. What is more, from this model we can see that *languages* influence positively in the amount of unique *views*, as the video is available for more people. Finally, *published_date* is positively correlated with the amount of *views* is that TED Talks have more viewers than the past who are more interested in recent topics rather than old videos.

Comments (strength and limitations)

1. This dataset lacks data necessary to analyze the monetization of views, for example, certain amount of views per published YouTube video. This additional data would allow our research to inform businesses in decisions having to do with profit.
2. The regression results display the model of best fit as model number 3, where *duration* and *views* have a positively correlated relationship. While this relationship is plausible, we believe that our dataset does not contain enough videos longer than 30 minutes to create a realistic interpretation of whether longer videos yield higher or lower unique *views*.
3. The *duration* only affects views until one point, 25 minutes, after this point other factors are the ones that affect the continuous increase of *views*. While *languages* and *published_date*, included in the model, increase views, there might be other factors which we cannot observe because they are not in the data set.

4. The advantage of our model is that it is based on a data set that has 2550 observations, which makes the model more accurate. What is more, all coefficients are significant at a 99 percent confidence level.

d. Conclusion

Managerial Implication / Recommendation

1. The dataset shows that 98 percent of the videos have a *duration* amount under 25 minutes (see plot Relation between Views and Duration). We can see that there is a positive correlation between *duration* and *views* until 25 minutes. Where duration is equal to more than 25 minutes, the number of views are no longer positively affected by the duration.
2. When the duration is very large, such as more than 50 minutes, we can see that there are less views. Therefore, we recommend that duration of a video should be no more than that 25 minutes for maximizing views.

What you learned from this project/analysis or Why this is useful

1. As our audience is video production companies, pinpointing the most influential variable (*duration, published_date, comments, or languages*) in determining unique *view* count, allows us to comment on what most drives views to youtube videos, which indirectly influences profit.

==

PROBLEM 2

A. Question: Does the type of tag affect the number of views ?

B. Background:

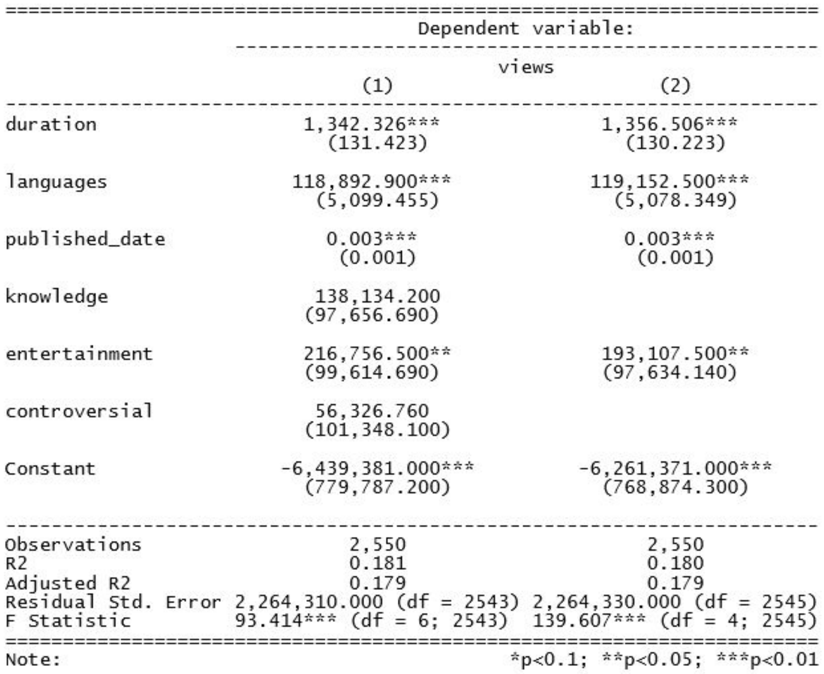
a. Previous researchers have observed that Youtube video watchtime is positively correlated with positive comments posted on Youtube videos (Yang et al., 2016). We hypothesize that the more controversial tags will have a higher view count present than non-controversial tags.

b. Additionally, since previous researchers have observed that a diversity of interest in relation to controversial tags stimulates social media users such as the most active Twitter users mentioning a video (Yu et al., 2014), this background evidence further supports this hypothesis.

C. Empirical Results & Discussion

a. Model 1: Multiple Linear Regression Model using binary variables

b. Model 2: Multiple Linear Regression Model using binary variables removing the independent variables *knowledge* and *controversial*.



```
===================================================================
                              Dependent variable:
                   ------------------------------------------------
                                        views
                           (1)                      (2)
-------------------------------------------------------------------
duration              1,342.326***             1,356.506***
                       (131.423)                (130.223)

languages            118,892.900***           119,152.500***
                      (5,099.455)              (5,078.349)

published_date          0.003***                 0.003***
                        (0.001)                  (0.001)

knowledge             138,134.200
                      (97,656.690)

entertainment         216,756.500**            193,107.500**
                      (99,614.690)             (97,634.140)

controversial          56,326.760
                      (101,348.100)

Constant            -6,439,381.000***        -6,261,371.000***
                      (779,787.200)            (768,874.300)

-------------------------------------------------------------------
Observations              2,550                    2,550
R2                        0.181                    0.180
Adjusted R2               0.179                    0.179
Residual Std. Error 2,264,310.000 (df = 2543) 2,264,330.000 (df = 2545)
F Statistic         93.414*** (df = 6; 2543)  139.607*** (df = 4; 2545)
===================================================================
Note:                                   *p<0.1; **p<0.05; ***p<0.01
```

c. Dataset
   i. *Views*: Measured in the amount of unique views per user
   ii. *Entertainment*: measures if a video includes tags such as comedy or humor
   iii. *Controversial*: measures if a video includes tags such as race or politics
   iv. *Knowledge*: measures if a video includes tags such as education or math

d. Results & Interpretation
   i. Model 1: We ran a Multiple Linear Regression Model using binary variables because, we are interested in analyzing how different types of topics influence *views*. We extended our model from question 1 to add three new binary variables (tags) related to topics of TED Talks. Prior to running this regression, we expected to see controversial tags cause higher number of views since people are more interested in watching controversial topics. After running the regression, we observe that the variable *controversial* is not statistically significant. Therefore, controversial topics are not shown to support increasing number of views from our regression model. However, the variable *entertainment* was shown to be statistically significant at a 95% confidence level.
   ii. Model 2: We decided to create another model without the independent variables *knowledge* and *controversial* because they are not significant. Running this model, we obtain that all the coefficients are statistically significant.
   iii. Chosen Model: TED Talks are popular for their specific formula where presenters try to introduce the topic interlacing it with humor and jokes,

because of that the reality is represented better by model (2) and it not enough to present just an interesting and knowledgeable topic. TED Talk videos, which present the topic in a humoristic way, become easier to understand, and as a result their number of views increase.
Thus, the model chosen for predicting the number of views for the question 2 is as follows:

$$views = \beta_0 + \beta_1 * duration + \beta_2 * languages + \beta_3 * published\_date + \beta_4 * entertainment$$

e. Comments (strength & limitation)

   i. We have to keep in mind that creating tags is very subjective and depends on the person. While for some people one video might be classified as controversial within a tag, for others that may not be the case.

D. Conclusion

Managerial Implication / Recommendation

1. While choosing a substantial topic is crucial, TED Talks presenters should not forget about the delivery of how they present their information. We recommend if their topic is educational, then they also include humor and jokes to make it easier for understanding, thus maximizing engagement.
2. A competing firm would benefit from seeing that the variable *entertainment* is the most reliable indicator in determining number of unique *views*. With this information the firm would be able to target the type of video to gain the highest number of unique *views* per video.

What you learned from this project/analysis or Why this is useful

1. These findings can be useful to potential sponsors who want to target the entertainment tags to monetize number of unique *views*. With this information, provided sponsors could decide which tags to advertise their products on for potential advertising revenue from those targeted videos.
2. Additionally, if speakers also want to maximize that engagement and potentially increase their viewership of their TED Talks, then implementing humor into their spoken deliveries would be beneficial.

==

PROBLEM 3

a. Question: How does the type of tags influence the publish date (vacation)?

b. Background:

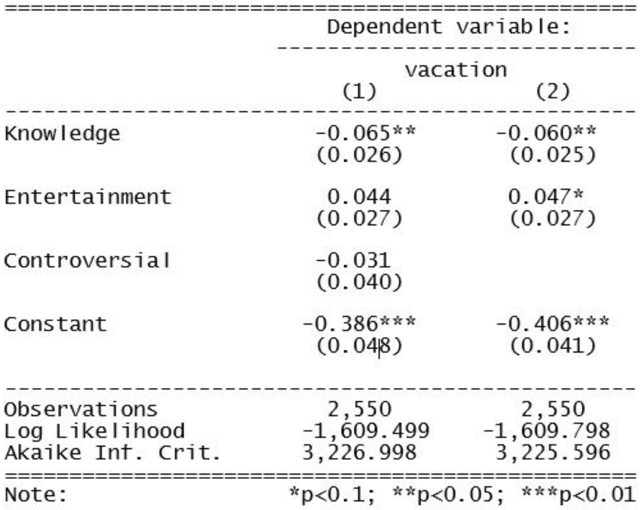
   i. Previous researchers have observed that more controversial tags attract the most comments and debate, (Sugimoto & Thelwall, 2013). We hypothesize that there are less Knowledge tags and more entertainment tags during vacation months. We have chosen to call the vacation months from May to August.

   ii. Additionally, since previous researchers have observed that Twitter-derived features through tags alone can predict whether a video will be in the top 5 percent in popularity (Yu et al., 2014), this background evidence further supports this hypothesis.

c. Empirical Results & Discussion

i. Model 1: Probit regression having *vacation* as dependent variable.

ii. Model 2: Probit regression deleting the independent variable *controversial*.



```
================================================
                      Dependent variable:
                  ------------------------------
                           vacation
                      (1)            (2)
------------------------------------------------
Knowledge          -0.065**       -0.060**
                   (0.026)        (0.025)

Entertainment       0.044          0.047*
                   (0.027)        (0.027)

Controversial      -0.031
                   (0.040)

Constant           -0.386***      -0.406***
                   (0.048)        (0.041)

------------------------------------------------
Observations        2,550          2,550
Log Likelihood    -1,609.499     -1,609.798
Akaike Inf. Crit.  3,226.998      3,225.596
================================================
Note:             *p<0.1; **p<0.05; ***p<0.01
```

iii. Dataset (linked to variables in the question)

1. *Vacation*: Measured as a binary variable where the vacation time frame is May through August. September through April is the non-vacation time frame.
2. *Knowledge*: measures the amount of educational tags included in a video such as education or math
3. *Entertainment*: measures the amount of entertaining topics in a video, such as comedy or humor
4. *Controversial*: measures the amount of controversial topics in a video, such as race or politics

iv. Results & Interpretation

1. Model 1: We ran a probit regression model because the variable that we wanted to analyze is binary. For the regression analysis we decided to use the probit model with *vacation* as the dependent variable and *knowledge, entertainment,* and *controversial* tags as the independent variables. Pr(*Vacation*=1|*knowledge*, *entertainment*, *controversial*) = $\Phi(\beta_0 + \beta_1$ **knowledge* + β_2 **entertainment* + β_3 **controversial*)
   Prior to running the regression model in problem 3, we expected to see the number of knowledge tags to decline during the vacation season due to the

amount of leisure time an individual has between the months of May through August. After running the regression, we can determine that our initial conclusion is valid due to the negative coefficient from the *knowledge* tag as well as the correlation between *entertainment* tags and *vacation* time. With this information, we can determine that an individual will more likely watch an entertaining video rather than a *knowledge* video during vacations.

2. Model 2: To further validate our claim, the controlled variable “*Controversial*” tags was added to the regression analysis to account for an omitted variable bias. Due to the coefficients slight change with the variable added, we can conclude that the *controversial* tag is not significant and does not affect the probability that a controversial tag would be on a vacation.
3. Chosen model: As a result, the variable *controversial* tag is not included in our final regression (model 2) analysis because the regression does not suffer from an omitted variable bias.

v. Comments (strength & limitation)

1. In order to determine whether the *published_date* is during a vacation period or not we set a binary variable to 1 for months May through August, which we consider that there is a vacation period. Our limitation for this measure was including every country in the world because different countries have different vacation months out of the year.
2. To solve the limitation we decided to use the northern hemisphere as our measurement. This would allow us to account for roughly 90% of the world's population due to the northern hemisphere having similar vacation seasons.

d. Conclusion

Managerial Implication / Recommendation

1. Based on the results, we recommend that during vacations the topic of the talk should not be about knowledge. We recommend TED Talks publish less educational topics during vacation months because their expectation is that during this period, viewers watch videos to have fun rather than with an academic purpose.
2. Instead, we recommend that the topic of the video would be more related with topics such as comedy and humor.

What you learned from this project/analysis or Why this is useful

1. These results can be useful for a new video publishing business that is interested in creating a similar product, to know when their competition publishes each type of video.

2. TED Talks speakers might also be interested in this finding, because if their topic is knowledgeable, the probability that their video would be published during vacations is low.

==

OVERALL SUMMARY

This information can be useful for companies that publish TED Talks videos associated with their business actions. For instance, knowing that TED Talks published with more entertaining topics during holiday breaks could be relevant information for a new company to decide what topic to publish during that time. Speakers also might be interested in these findings in order to increase their audience engagement and increase the viewership of their TED Talks videos. Additionally, Ted presenters, company sponsors, and video publishers can take into account these findings in relation to the optimal durations of TED Talks videos and the topics they cover.

REFERENCES

1. Banik, Rounak. “TED Talks Dataset .” India, 2017. https://www.kaggle.com/rounakbanik/ted-talks
2. Park M, Naaman M, Berger J. (2016). A Data-Drive Study of View Duration on Youtube. *Proceedings of the Tenth ICWSM 2016.* http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/download/13062/12820.
3. Guo PJ, Juho K, Rubin Rob. (2014). How Video Production Affects Student Engagement: An Empirical Study of MOOC Videos. *Proceedings of the First ACM Conference on Learning 2014.* https://dl.acm.org/citation.cfm?id=2566239.
4. Yang R, Singh S, Cao P, Chi E, Fu B . (2016). Video Watch Time And Comment Sentiment: Experience From YouTube. *2016 Fourth IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb).* https://ieeexplore.ieee.org/abstract/document/7785813.
5. Yu H, Xie L, Sanner S. (2014). Twitter-Driven YouTube Views: Beyond Individual Influencers. *Proceedings of the 33nd ACM International Conference on Multimedia 2014.* https://dl.acm.org/citation.cfm?id=2655037.
6. Sugimoto CR & Thelwall M. (2013). Scholars On Soap Boxes: Science Communication And Dissemination In TED Videos. *Journal of the American Society for Information Science and Techniology.* Vol. 64, Issue 4, February 2013. https://onlinelibrary.wiley.com/doi/full/10.1002/asi.22764.
7. Boppoloige AA & Gurtoo A. (2017). What Determines Viral Phenomenon? Views, Comments, And Growth Indicators of TED Talk Videos. *International Journal of Trade, Economics and Finance.* Vol. 8, No. 2, April 2017. http://www.ijtef.org/vol8/544-EM0013.pdf.