Analysis of Influence Indicators of YouTube Videos using Metadata

Robert Brouwer
Perseusstraat 23 7521ZA Enschede
The Netherlands
r.j.brouwer@student.utwente.nl

ABSTRACT

Youtube is one of the biggest social media platforms. It is used to share and view video's. This research looks at which video metadata are potential indicators for influence. For this four different metrics are used that highlight different aspects of influence. Metrics used are: percentage of likes of total likes & dislikes, likes & dislikes per view, comments per view and views per subscription. Video category is found to be the most decisive aspect creating distinction between metrics. Video length and view count show weak influence. Video count and subscriptions show limited influence.

Keywords

YouTube, Video Metadata, Influence Indicators, Correlation Analysis

1. INTRODUCTION

Social networks are a common part of everyday life. They influence everyday activities and are one of the main sources for news. Users include people from all parts of society including influencers, politicians and vloggers, but also many people with no specific interests. This study focuses on one specific social network, YouTube.

YouTube is a platform which can be used to view and share video's. With about a quarter of the earth using the platform and about 500 hours of video uploaded every minute[4] it is one of the biggest social networks there is. Being so big it provides people with information about various topics. People are informed about video's via recommendations through the YouTube algorithm, via links from other platforms, or via subscription to a channel.

Due to the size of the platform it has influence on the users. Absolute influence of one video is nearly impossible to quantify. Judging the difference in influence between videos may be assessed through observing indicators based on metadata.

Indicators of influence that might be applicable are:

- Percentages of likes of total likes and dislikes (Rating)
- Likes & dislikes per view (Light engagement)
- Comments per view (High engagement)
- Views per subscription (Relative reach)

The rating is an indicator of how aligned a video is with the audience. Likes & dislikes per view and comments per view are both indicators for engagement. This shows what percentage of the viewers that took the effort to press the like/dislike button or write a comment. Views per subscription is an indicator for reach compared to the build audience.

This study researches the correlation between the above metrics for influence and different video metadata considering this as potential predictors for influence.

31th Twente Student Conference on IT Jul. 5nd, 2019, Enschede, The Netherlands.

2. RELATED WORK

One of the valuable resources of YouTube is it's comment section. Collecting opinion about vaccination has been done by Yiannakoulias, Slavik & Chase[10] with the use of the comment section. Retrieving opinion over a video has also been done by Thelwall[8] using the comments of the video. The comment section of a video, when analyzed using sentiment analysis, gives good insight in people's opinion about a subject. These studies show that the comment section is a good source for retrieving the sentiment of the audience. Therefor comments might also be a good measure for the influence.

Hoiles, Aprem & Krishnamurthy[5] have researched what makes a video receive more views. They found that data such as first day view count, number of subscribers and contrast in thumbnail among other things can predict the view count. One of the main focuses in this research is the number of subscribers. For this they show that more subscribers result in more views and vice versa. However this study doesn't show details on this relationship.

There are many different possible reasons why somebody might leave a like or dislike on a video. Shoufan[6] has done an exploration for reasons why students might like or dislike an educational video. Main reasons for liking/disliking included aspects as explanation & understanding, presentation methods and content. Within this study 12% of the students included length as a reason for liking/disliking the video.

The way that information is delivered influences how people like/dislike a video. Djerf-Pierre, Lindgren & Budinski[3] compared the difference in engagement on videos about antimicrobial resistance. They compare videos that where defined as popular science to those defined as journalism. They found that popular science videos had on average more views, more likes per view and less dislikes per view. Comments per view didn't show any difference

in performance. Although subject and message is almost the same, deliverance highly influences how people interact.

3. DATA COLLECTION

Test data has been collected through the use of the YouTube Data API V3[2]. This can be used to directly collect the statistics of videos and channels from YouTube. Calling the YouTube API[2] and saving data has been done using Python[9]. Table 1 shows the variables that are collected.

Table 1. Data collected about videos and channels

Video	Channel
Video ID	Channel ID
Title	Channel Name
Publication Date	Publication Date
Description	Language
Category	Total Views
Language	Total Comments
Length	Subscriptions
Views	Number of Videos
Likes Count	
Dislikes Count	
Comments	

Collection of the data involved three major parts: ID collection, Data retrieval and Filtering

3.1 ID Collection

To gather the statistics of an video the ID is required. Lists of video ID's with the right requirements are not publicly available, thus has to be created. This is done by using the search function of the Youtube API[2]. By using a query it returns all video ID's that fit the request.

3.1.1 Filters

To only return videos of interest the YouTube API[2] has the possibility for filters.

As upload period March 2019 was used whereas data collection been done second half of May 2019. This period was selected to be far enough back for videos to gain views, but recent enough to be still relevant for the platform. SzabÃş and Huberman[7] have looked at how view counts develop over time. They couldn't show a moment videos stop gaining views, so growth is expected to be indefinitely. So observing number of views is by definition a snapshot.

Originally the area of the Netherlands was chosen as the location, but this was later removed. The problem was that only videos which had an location tag where returned. This is however not the case for most videos and would limit the data set. Most videos returned from this where from festivals and tourist attractions.

Filtering on language was set to Dutch. However this filter was removed due to it performing very poorly. Videos which contained other languages such as Russian, Hebrew or Italian would not be filtered out. The YouTube uses the language filter only as an suggestion not as a hard requirement. Videos which do fit the rest of the criteria, but not the language will still be returned. Due to it's unreliability this filter was dropped.

3.1.2 Randomness

To limit the possible biases in the data it is preferred to collect videos randomly. However true random collection of data was unfortunately not possible. Two methods which in theory would return random values where tested, but both methods have their own limitations.

The video ID is a pseudo random string of eleven characters. Randomly creating a string that matches an existing ID is impossible. Online suggestions on using part of the string to return a video also doesn't work. When searching for part of a random string this returns videos that have this sub string contained in their video description. Only limited videos have a link in their description, so this method is very biased.

Another possible method of selecting a video randomly would be by using the upload moment. This method would work by selecting a certain time stamp and order the videos on upload moment. This method unfortunately also doesn't work. Although it is possible to order videos on date the returned video is not the one latest uploaded. When a query is only based on date the returned list of videos is a selection which is created using the preferences of the platform.

The method used in this study is by using as random search query of one or two letters at a randomly selected day in March 2019. The first fifty returned ID's are added to the data set. Fifty is the maximum amount the YouTube API[2] allows to be returned per call. The one or two letters are chosen with a uniform distribution. This method is biased towards videos that contain in the title or description words with less common letter combinations. However by using a small query is it limited to the minimum. Also their is a bias towards videos that YouTube ranks higher.

3.2 Data Retrieval

The list of video IDs created is first cleaned from duplicate IDs. In the way the IDs are collected it can happen that two different queries might return some of the same values.

Using the YouTube API[2] the statistics of the videos are collected. This also gives back the channel ID which can be used to collect the statistics of the channel. Because the YouTube API[2] works with a quota system. This limits the collection of video statistics to 500 videos per day per account. In total 14,233 different videos are gathered.

Some of the text returned contained characters are not part of the European alphabet. To prevent errors that might occur from this these characters got removed.

3.3 Filtering

To make sure that all data used in the analysis is also useful there are a few requirements made.

The data point must be complete. Some of the video ID's retrieved do not have video metrics or do not have a corresponding channel. Lack of this data can be due to privacy settings of the video or removal of the video or channel. The real reason behind why the data is not available is not always clear.

Videos must have at least 1,000 views. This is to make sure that the video has seen at least enough traffic to make values as amount of comments per view sensible. When the view count is very low the influence one user has on these values becomes very high, meaning the values are based on the actions of individual users instead of the group. After the filtering their where in total 8,889 unique data points.

4. RESULTS

4.1 Terminology and measures

To calculate the values and creation of corresponding graphs MATLAB[1] was used. The results are quantified using different well known statistical methods.

The Pearson coefficient is an indicator of how close the values are to the best fit linear relationship. The value varies between -1 and 1 where values close to 1 indicate a strong positive correlation and -1 a strong negative correlation. Values close to 0 mean that there is no correlation.

The Spearman coefficient indicates if there is a correlation based on the order of the values. This value does not care about the shape of the correlation, but gives an indication about the direction of the correlation and its strength. It measures how monotonic a certain relationship is. A relation is monotonic if it's always increasing or decreasing. The value stretches from -1 to 1 where negative values show a negative correlation and positive values a positive. Values close to 1 or -1 show a strong correlation and values close to 0 indicate no correlation.

For tests comparing categories the unequal variance t-test is used. This test is selected because there is no common variance that applies to all categories. The t-test shows if the mean of two distributions statistically significant differ. This study applies the T-test to compare the mean of any category to the rest of the data. Since it is not a test between two categories it does not tell if one category performs better than another. It only tells if the category performs significantly better or worse than the mean. The one-sided test is used since the distributions seem to follow a right skewed curve. Outliers are not used in the calculated t-test. This to avoid the extreme influence they have on the average of an metric and creating values not reflecting the rest of the category. Values are considered as outlier when it is more than 1.5 times the difference between the 25 and 75 percentile.

The p-value shows if a correlation is statistically significant. This research uses a significance level of 5%. This means that any coefficient which has a p-value below 0.05 is seen as statistically significant. It is the likelihood of rejecting a correlation when it does exist.

Also the power of the T-test is calculated. This value shows how likely it is that the assumption that the difference exist is indeed true. In the t-test tables the degrees of freedom (DF) and t-score are added for completeness.

Table 2. Category's number and name as classified by YouTube[2]

Nr.	Category Name	Nr.	Category Name
1	Film & Animation	23	Comedy
2	Auto & Vehicles	24	Entertainment
10	Music	25	News & Politics
15	Pets & Animals	26	Howto & Style
17	Sports	27	Education
19	Travels & Events	28	Science & Technology
20	Gaming	29	Nonprofits
22	People & Blogs		

Table 2 lists the number of every category and the corresponding description. The names of the categories and the corresponding number are the same as assigned by YouTube and can be found through using the API[2]. This list only contains the category's that are found in the created data set. Other category's might not be in use anymore or are rarely assigned.

4.2 View Count

Table 3 shows the correlation coefficient between the three metrics and the amount of views.

Table 3. Correlation coefficients and p-values between Amount of Views and influence indicators

	Rating	Likes/Dislikes per view	Comments per view
Pearson Rho	0.00115	-0.0868	-0.0659
P-value	0.914	2.32e-16	4.82e-10
Spearman Rho	-0.0120	-0.149	-0.298
P-value	0.261	2.71e-45	2,52e-181

The rating does not give any indication of a correlation although it would seem logical that such a relationship would exist. This because if a larger proportion likes a video than it would be expected that more people would watch it. However, both metrics are not able to give a reliable indication.

Both the Pearson and Spearman coefficients show low results on the correlation between views and likes/dislikes per view. Since they are both so low it's assumed that there exists no relationship between them. However it does not mean that there might exist any other relationship. Such a relationship seems however unlikely because it should be non-monotonic.

For the comments per view no linear correlation is found, but there is a weak monotonic relationship. This can indicate that there is a good chance there exist a higher order correlation between views and comments per view. The relationship seems to exist but is very weak. An explanation for this might be that people are less likely to comment seeing there are already many comments. However to check this over time development of the comment-section of a video is needed. This time dependent data is however not collected and can not be tested within this research.

4.3 Video Length

Table 4 shows the relationship between the three metrics and the video length. Although difficult to spot with the naked eye there is a small tendency for having a lower rating when the length of the video increases (see figure 1). The rating decreases very slowly at about 1 percentage by an increase in time of 1 hour and 17 minutes. Both correlation factors show that although video length is an indicator it is not a very strong one, but does have influence on the rating. This corresponds with the outcome of Shoufan[6] which showed that video length is a reason for liking/disliking a video.

Table 4. Correlation coefficients and p-values between Video Length and influence indicators

	Rating	Likes/Dislikes per view	Comments per view
Pearson Rho	-0.0285	-0.0147	-0.00811
P-value	0.00723	0.167	0.445
Spearman Rho	-0.133	0.110	0.223
P-value	2.80e-36	2.203e-25	7.914e-101

Although there is no indication of any linear correlation there is a tendency for people to press the like/dislike button more with increasing length. It is a weak correlation, but it shows length has influence on the interaction with the video. The same relationship exist with the comments, but this relationship is stronger than with the like/dislikes.

It is still a correlation which would be described as weak, but of the three indicators the correlation with comments is strongest. Reasons for why the correlation is stronger with comments than with likes and dislikes might be that creation of a comment costs more creativity. Having a longer video gives more time for items of interest to come along and therefor to comment.

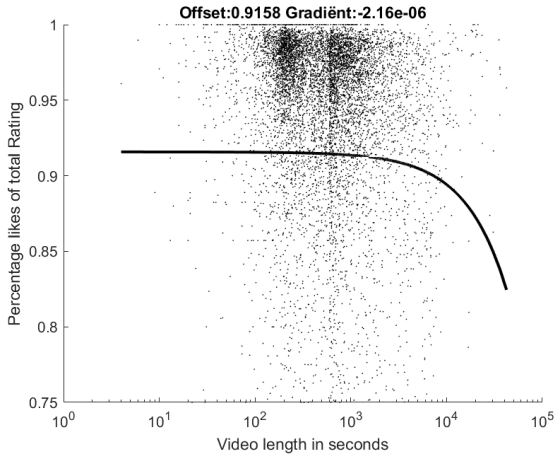

Figure 1. Scatterplot with linear best fit line between Rating and Length of video

4.4 Category

This section considers the relationship between the category a video belongs to and how people interact with it when it comes down to pressing the like/dislike button or writing a comment.

4.4.1 Rating

When looking at the rating of a video one of the categories which is most different from the rest is News & Politics(25). Having an mean of about 90% which is far lower than the overall as can be seen in table5. The overall mean is close to 95%. The difference can easily been seen in figure2. This makes sense since this category produces content that has opinions incorporated in it. When people do not agree with the opinion of the video they might be more inclined to leave a negative review.

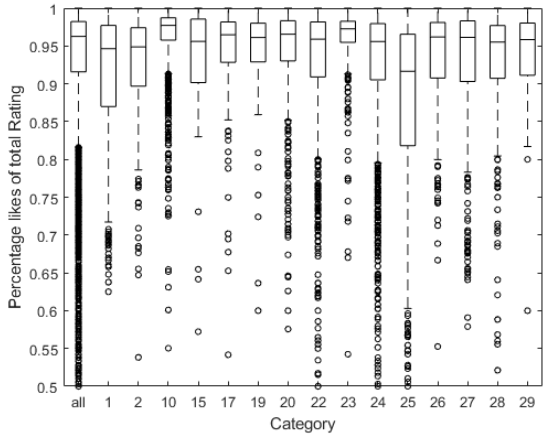

Figure 2. Boxplot with Rating distribution per Category

Categories with a high rating are Music(10) and Comedy(23). Both are relatively light entertainment and have a mean rating of about 97%. Both are apparently the type of entertainment which is liked by many people.

Table 5. One-sided unequal variance t-test of Rating done by Category

	t-score	DF	p-value	Power	Nr.	Mean
All	-	-	-	-	7957	0.954
1	6.86	328	1.69e-11	1.00	320	0.927
2	5.00	297	4.87e-7	1.00	282	0.941
10	34.0	5384	0.00	1.00	1498	0.975
15	0.184	7955	0.427	0.0720	23	0.956
17	0.801	259	0.212	0.198	240	0.956
19	1.05	59.4	0.150	0.267	59	0.959
20	4.38	709	6.96e-6	0.997	564	0.960
22	3.10	1432	0.00101	0.926	1125	0.950
23	15.4	507	0.00	1.00	343	0.971
24	7.45	2543	6.26e-14	1.00	1760	0.947
25	15.7	639	0.00	1.00	621	0.897
26	2.50	363	0.00643	0.802	342	0.950
27	0.370	445	0.356	0.101	413	0.953
28	3.54	7957	0.000203	0.971	371	0.947
29	0.912	7955	0.181	0.232	72	0.950

4.4.2 Likes/dislikes per view

Figure 3 shows the amount of likes/dislikes as fraction of the views. As worst performing category is Auto & Vehicles(2). This is followed with a small gap with Sports(17) and Travel & Events(19) (See table 6). Expected reasons why they perform lower than other categories is that they don't leave a high impact on the viewer, thus being less likely to press any of the buttons.

Education(27), Howto & Style(26) and Comedy(23) perform the best. There is no direct connection between these categories. Especially since some categories are closer to other categories like Education(27) and Science & Technology(28) that have a clear gap in performance. Djerf-Pierre, Lindgren & Budinski[3] already showed that videos which lay close together when it comes to content might not perform similarly.

Table 6. One-sided unequal variance t-test of Likes/Dislikes per View done by Category

	t-score	DF	p-value	Power	Nr.	Mean
All	-	-	-	-	8421	0.0225
1	5.68	8416	6.91e-9	1.00	335	0.0173
2	13.7	353	0.00	1.00	282	0.0147
10	2.47	2756	0.00671	0.796	1581	0.0216
15	0.505	26.08	0.309	0.120	27	0.0249
17	12.4	279	0.00	1.00	238	0.0143
19	2.07	8419	0.0192	0.664	63	0.0181
20	7.12	8421	3.28e-6	0.998	597	0.0256
22	3.67	1455	0.000127	0.978	1163	0.0246
23	5.43	8424	2.93e-8	1.00	353	0.0274
24	3.11	8416	0.000943	0.928	1851	0.0214
25	3.93	8417	4.31e-5	0.989	648	0.0200
26	5.14	379	2.17e-7	1.00	357	0.0279
27	7.08	497	2.45e-12	1.00	465	0.0295
28	3.01	382	0.00130	0.914	382	0.0251
29	0.168	8419	0.433	0.0698	76	0.0228

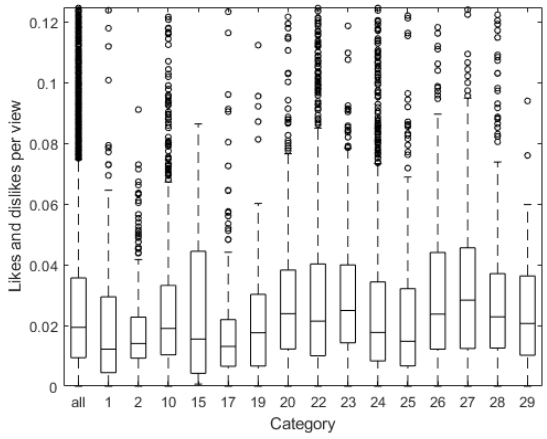

Figure 3. Boxplot with amount of likes and dislikes per view per category

News & Politics(25) would be expected to have a higher likes and dislikes value since it is a category which is based on opinions. However there performance based on the amount of likes/dislikes is below the overall average. This might be an indication that interaction within this category requires more energy compared to other categories, because the viewer needs to choose side.

4.4.3 Comments per view

Since comments take more effort to create a difference in which categories gain the most engagement might be visible. As very worst performer the category Music(10) stands out as seen in figure 4 and table **??**. When compared to it's performance with the likes and dislikes per view (see table 6 and figure 3) it performs far worse. This shows that there is a clear difference in how people use likes/dislikes and comments. A reason why music performs lower might be that there is just not much information to react to. Film & Animation(1) performs almost identical to Music(10). Since Film & Animation(1) contains mainly trailers the comparison to music clips is easily made.

As top performers are News & Politics(25) and Science & Technology(28). Both are information heavy categories which corresponds to why Music(10) might perform lower than the other categories.

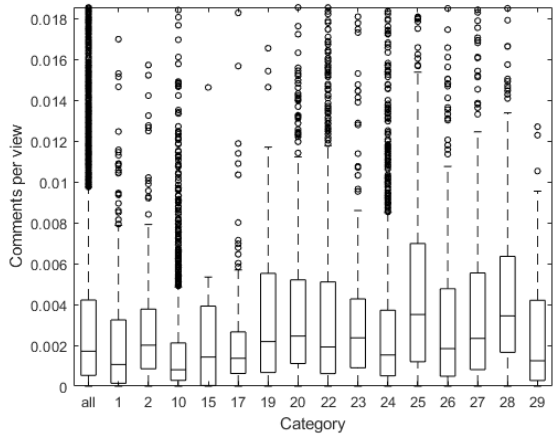

Figure 4. Boxplot with amount of comments per view per category

Table 7. One-sided unequal variance t-test of Comments per View done by Category

	t-score	DF	p-value	Power	Nr.	Mean
All	-	-	-	-	8145	0.0022
1	7.13	356	2.81e-12	1.00	317	0.0015
2	0.523	320	0.301	0.130	288	0.0023
10	32.4	4771	0	1.00	1490	0.0011
15	0.957	8143	0.169	0.246	25	0.0018
17	7.35	285	1.04e-12	1.00	241	0.0016
19	2.05	63.5	0.0225	0.646	64	0.0031
20	7.12	641	1.49e-12	1.00	573	0.0030
22	4.06	1383	2.56e-5	0.992	1130	0.0026
23	2.52	389	0.00612	0.807	349	0.0025
24	5.09	3342	1.86e-7	1.00	1803	0.0020
25	13.0	673	0.00	1.00	635	0.0041
26	1.28	354	0.101	0.355	332	0.0024
27	5.27	467	1.04e-7	1.00	439	0.0030
28	10.9	404	0.00	1.00	385	0.0039
29	0.448	8143	0.327	0.116	74	0.0021

4.4.4 Views per Subscription

Figure 5 and table 8 show the influence of category on the views per subscriber.

Music(10) has a far higher mean than any of the other categories. This might be due to people watching the video multiple times, continuously as background music or due to people quickly searching for the content. Another possibility might be due to people illegally uploading music videos. Thus people will not subscribe due to it not being the creator.

Travels & Events(19) and People & Blogs(22) both perform also rather well. The content of the two categories are comparable. Why these two perform even far better than Howto & Style(26), which has a comparable type of content, is unknown.

Pets and Animals(15) also performs very good, but due to it's low frequency no conclusions can be taken from it.

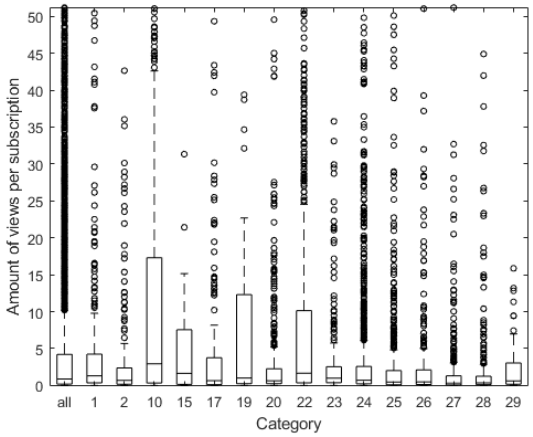

Figure 5. Boxplot with amount of views per subscriber as observed per category

Video categories that perform badly when it comes to views per subscriber are Education(27) and Science & Technology(28). Both categories are very close to each other when it comes to subject and performance. Reasons why they perform so badly might be that people are only interested in a small part of the content uploaded. In addition people may tend to subscribe to these kind of channels more due to it considered a valuable resource.

Table 8. One-sided unequal variance t-test Views per Subscription done by Category

	t-score	DF	p-value	Power	Qty	Mean
All	-	-	-	-	7450	1.44
1	2.32	7448	0.0103	0.749	296	1.71
2	5.00	320	4.78e-7	1.00	268	1.03
10	19.2	1444	0.00	1.00	1418	6.08
15	1.59	22.0	0.0627	0.451	23	2.712
17	1.12	7448	0.131	0.301	219	1.28
19	2.26	57.1	0.0138	0.721	58	3.03
20	10.7	873	0.00	1.00	530	0.892
22	10.9	1090	0.00	1.00	1033	3.05
23	4.04	408	3.13e-5	0.99	320	1.144
24	10.6	4589	0.00	1.00	1684	1.022
25	14.5	1013	0.00	1.00	569	0.743
26	9.03	431	0.00	1.00	322	0.852
27	22.7	946	0.00	1.00	411	0.510
28	21.7	802	0.00	1.00	356	0.545
29	0.0960	7445	0.462	0.0607	71	1.41

4.5 Subscriptions and video count

A channel changes in characteristics when it growths in size. For this the amount of subscriptions is compared to the views per subscription. When looking at figure 6 there is a clear downwards trend visible. This is affirmed by the Spearman coefficient as found in table9. The bottom left corner of figure6 is empty.This is due to the filtering of the data to not include any video's below 1,000 views. Correlation between subscriber count and amount of views per subscription might be less strong than it appears due to the created bias. No conclusions should be drawn from the empty left bottom corner.

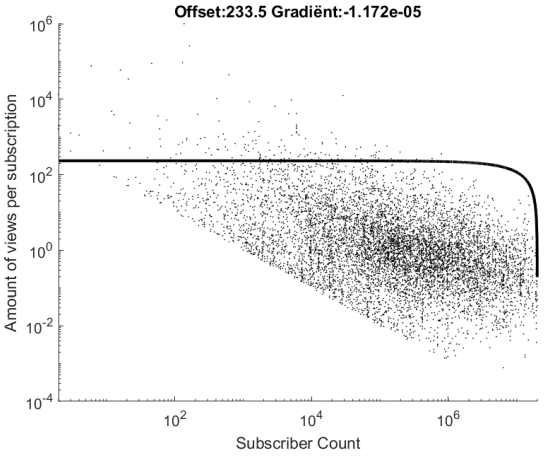

Figure 6. Scatterplot with linear best fit line between Subscriber Count and Views per Subscriber

Table 9. Correlation coefficients and p-values between Views per Subscription and different indicators

	Subscriptions	Video Count
Pearson Rho	-0.00586	-0.00421
P-value	0.723	0.580
Spearman Rho	-0.314	-0.249
P-value	3.67e-202	6.05e-126

Figure 7 shows that the amount of video's uploaded is correlated with the views per subscriber. Considering the correlation between video count and subscriber count as seen in figure8 the bias found in figure6 may also have some effect on this result. Considering the high certainty of the found correlation as shown in table 9 this limited bias does not change the conclusion.

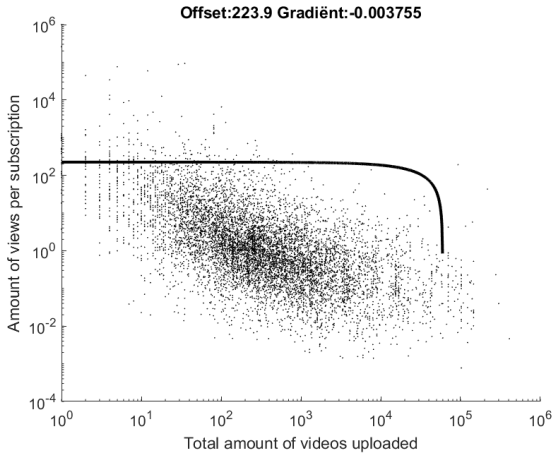

Figure 7. Scatterplot with linear best fit line between Amount of Videos and Views per Subscription

Table 10 shows a very strong correlation between the amount of videos uploaded and the amount of subscriptions. This correlation is also clearly visible in figure 8. Although the linear correlation doesn't fit very good, the gradient of the best fit linear relationship is already steep. This points out that the number of uploads highly increases the amount of subscriptions. For now an increase of 63 subscriptions with every video upload is observed.

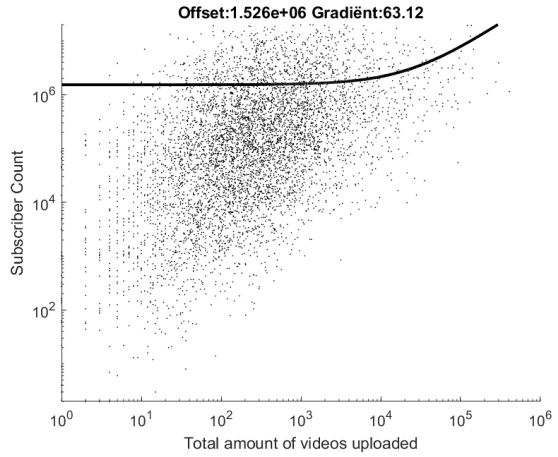

Figure 8. Scatterplot with linear best fit line between Video Count and Subscriber Count

Table 10. Correlation coefficients and p-values between Views per Subscription and Video Count

	Video Count
Pearson Rho	0.1414
P-value	6.03e-41
Spearman Rho	0.4991
P-value	0

It is expected that this relationship works in both ways. It could be said that channels that upload more have a bigger chance on creating new subscriptions due to having more content. However the other way around could also easily be explained. Because they have more subscriptions they are more encouraged and gain more resources to create content.

5. CONCLUSION

This study tried to find the potential of predicting influence through observation of video metadata. The study considered four different metrics to measure influence of a video. All observed metadata showed a basic level of predictability. The study observed that category is of major impact on statistics of all metrics. Length and view count show weak influence. Video count and subscriptions show limited influence. All metadata is by itself not able to predict influence combining them may lead to a possible good predictor. Due to the choice to only use videos with at least 1,000 views their might be a bias. This bias mostly influences results concerning subscriber count and videos uploaded.

6. DISCUSSION

Because the YouTube API[2] was used to gather video data there is a bias created by how YouTube selects. There is no way to exactly determine how you want videos selected or how the used algorithm works. It is difficult to quantify the bias, however in future research this could be minimized further. This could happen by highly increasing the amount of videos collected or by circumventing the algorithm.

As seen in figure 6 filtering on views can create a bias. However since videos with low amount of views have a high variance filtering is necessary given the methods used within this study. When videos with low amounts of views are left in values such as comments per view can become unrealistically high or low. Results of such study could be enhanced by defining a method in which videos with low amounts of views can be incorporated while still creating statistically significant results.

During the test and when discussing the results it's assumed that the categories where correctly assigned. However when exploring some of the data it was observed that not everything had been assigned the category that might be expected. A common example would be videos which are clearly vlogs being assigned to Entertainment(24) instead of People & Blogs(22). Quantifying the significance of these mistakes is difficult and would require manual checking of videos on a large scale.

To predict a correlation between two variables two different indicators are used. Linear correlation is most of the time not a good indicator. Possibly higher polynomial functions would gain better results and could be explored in further research. It would be realistic to expect a stronger correlation and make for a better indicator. However big improvement may be unrealistic due to how scattered the data is.

All indicators explored where directly measurable values. Further research could improve by using the channel's activities and derived values from the thumbnail. By including more information about the channel there could be gained insight in the exact role of the creator for engagement. Using the thumbnail could show more how the viewer is influenced by the image.

7. REFERENCES

[1] MATLAB and Statistics Toolbox Release 2012b, The MathWorks, Inc., Natick, Massachusetts, United States.

[2] G. Developers. Data api. URL:https://developers.google.com/youtube/v3/ Accessed on:11-5-2019.

[3] M. Djerf-Pierre, M. Lindgren, and M. A. Budinski. The role of journalism on youtube: Audience engagement with 'superbug' reporting. *Media and Communication*, 7(1):235–247, 2019.

[4] DMR. 160 youtube statistics and facts(2019)|by the numbers. URL:https://expandedramblings.com/index.php/youtube-statistics/ Accessed on 10-5-2019.

[5] W. Hoiles, A. Aprem, and V. Krishnamurthy. Engagement and popularity dynamics of youtube videos and sensitivity to meta-data. 29, July 2017.

[6] A. Shoufan. What motivates university students to like or dislike an educational online video? a sentimental framework. *Computers and Education*, 134:132–144, 2019.

[7] G. SzabÃş and B. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53, 12 2008.

[8] M. Thelwall. Social media analytics for youttube comments: potential and limitations. 21, September 2017.

[9] G. Van Rossum and F. L. Drake Jr. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

[10] N. Yiannakoulias, C. E. Slavik, and M. Chase. Expressions of pro- and anti-vaccine sentiment on youtube. 37, November 2018.

APPENDIX

A. BOXPLOTS

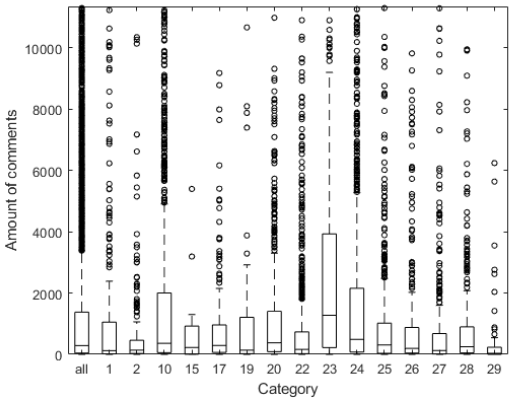

Figure 9. Boxplot with Amount of Comments per Category

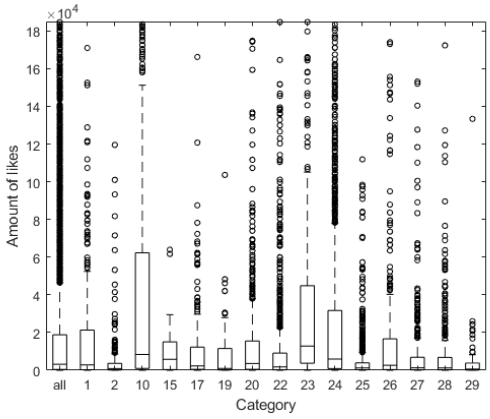

Figure 10. Boxplot with Amount of Likes per Category

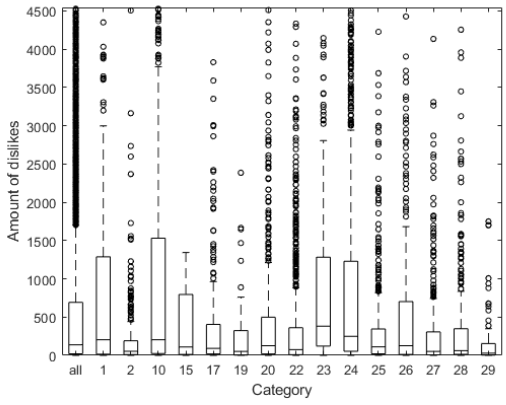

Figure 11. Boxplot with Amount of DisLikes per Category

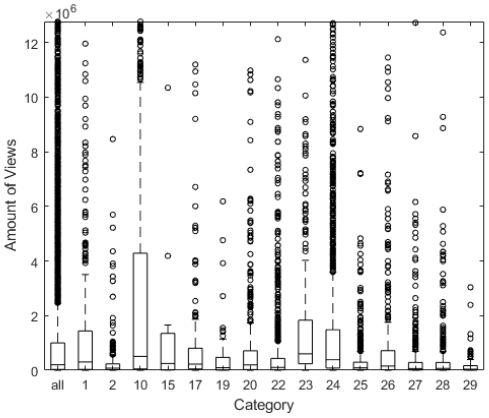

Figure 12. Boxplot with Amount of Views per Category

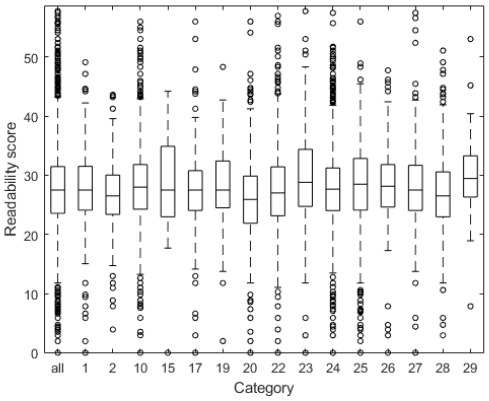

Figure 13. Boxplot with Readability per Category

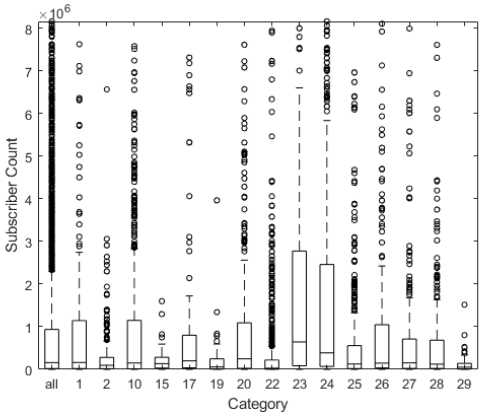

Figure 14. Boxplot with amount of Subscriber Count per Category

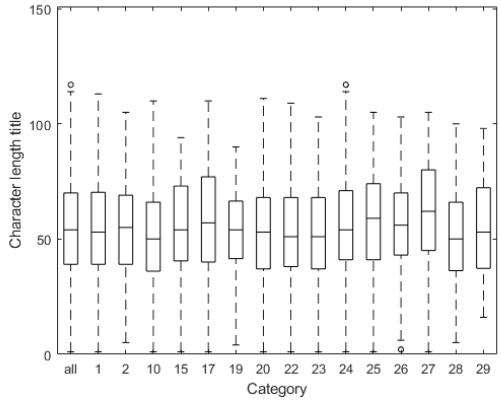

Figure 15. Boxplot with Title Length per Category

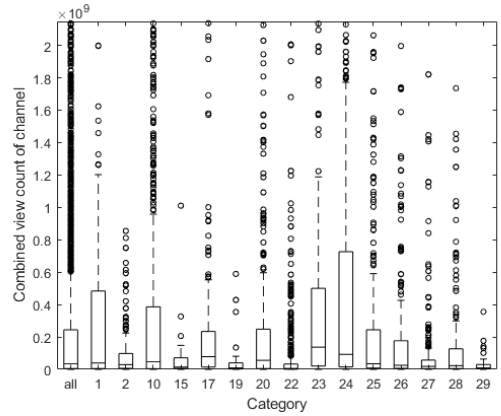

Figure 16. Boxplot with Total Views of a Channel per Category

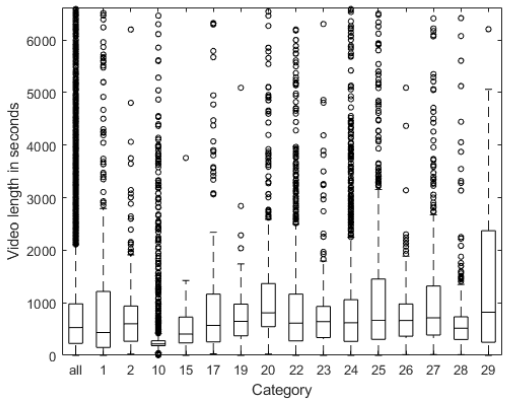

Figure 17. Boxplot with the Length of a Video per Category

B. CORRELATION VALUES

Table 11. Pearson correlation values part 1

	Views	Likes	Dislike	Comment	Rating	Title Len.	Com p. View
Views	1.0000	0.7190	0.7670	0.4400	0.0011	-0.0286	-0.0660
Likes	0.7190	1.0000	0.6580	0.8060	0.0660	-0.0489	-0.0360
Dislike	0.7670	0.6580	1.0000	0.5050	-0.0550	-0.0335	-0.0369
Comment	0.4400	0.8060	0.5050	1.0000	0.0476	-0.0377	0.0432
Rating	0.0011	0.0660	-0.0550	0.0476	1.0000	-0.0292	0.0253
Title Len.	-0.0286	-0.0489	-0.0335	-0.0377	-0.0292	1.0000	-0.0022
Com p. View	-0.0660	-0.0360	-0.0369	0.0432	0.0253	-0.0022	1.0000
(Dis)like p. View	-0.0869	0.0615	-0.0386	0.0735	0.2880	-0.0489	0.3920
Readability	0.0157	0.0419	-0.0004	0.0336	-0.0063	0.2510	0.0049
Tot Views	0.2040	0.1660	0.1810	0.1190	0.0135	0.0442	-0.0481
Subscr.	0.2280	0.2720	0.2240	0.2380	0.0399	0.0306	-0.0418
Tot videos	-0.0084	-0.0226	-0.0127	-0.0106	-0.0512	0.0128	-0.0124
Video Length	-0.0422	-0.0492	-0.0263	-0.0265	-0.0285	0.0245	-0.0081
View p. Sub	0.1240	0.1140	0.0761	0.0434	0.0039	0.0085	-0.0076

Table 12. Pearson correlation values part 2

	(Dis)like p. View	Readability	Tot Views	Subscr.	Tot videos	Video Length	View p. Sub
Views	-0.0869	0.0157	0.2040	0.2280	-0.0084	-0.0422	0.1240
Likes	0.0615	0.0419	0.1660	0.2720	-0.0226	-0.0492	0.1140
Dislike	-0.0386	-0.0004	0.1810	0.2240	-0.0127	-0.0263	0.0761
Comment	0.0735	0.0336	0.1190	0.2380	-0.0106	-0.0265	0.0434
Rating	0.2880	-0.0063	0.0135	0.0399	-0.0512	-0.0285	0.0039
Title Len.	-0.0489	0.2510	0.0442	0.0306	0.0128	0.0245	0.0085
Com p. View	0.3920	0.0049	-0.0481	-0.0418	-0.0124	-0.0081	-0.0076
(Dis)like p. View	1.0000	-0.0126	-0.0300	0.0432	-0.0895	-0.0147	-0.0083
Readability	-0.0126	1.0000	0.0119	0.0119	-0.0005	-0.0196	0.0063
Tot Views	-0.0300	0.0119	1.0000	0.8790	0.2080	-0.0369	-0.0038
Subscr.	0.0432	0.0119	0.8790	1.0000	0.1410	-0.0386	-0.0059
Tot videos	-0.0895	-0.0005	0.2080	0.1410	1.0000	-0.0082	-0.0042
Video Length	-0.0147	-0.0196	-0.0369	-0.0386	-0.0082	1.0000	-0.0074
View p. Sub	-0.0083	0.0063	-0.0038	-0.0059	-0.0042	-0.0074	1.0000

Table 13. P-values of Pearson correlation part 1

	Views	Likes	Dislike	Comment	Rating	Title Len.	Com p. View
Views	1.0000	0	0	0	0.9140	0.0070	0.0000
Likes	0	1.0000	0	0	0.0000	0.0000	0.0007
Dislike	0	0	1.0000	0	0.0000	0.0016	0.0005
Comment	0	0	0	1.0000	0.0000	0.0004	0.0000
Rating	0.9140	0.0000	0.0000	0.0000	1.0000	0.0059	0.0169
Title Len.	0.0070	0.0000	0.0016	0.0004	0.0059	1.0000	0.8340
Com p. View	0.0000	0.0007	0.0005	0.0000	0.0169	0.8340	1.0000
(Dis)like p. View	0.0000	0.0000	0.0003	0.0000	0.0000	0.0000	0
Readability	0.1380	0.0001	0.9680	0.0015	0.5500	0.0000	0.6420
Tot Views	0.0000	0.0000	0.0000	0.0000	0.2040	0.0000	0.0000
Subscr.	0.0000	0.0000	0.0000	0.0000	0.0002	0.0039	0.0001
Tot videos	0.4290	0.0328	0.2330	0.3160	0.0000	0.2280	0.2410
Video Length	0.0001	0.0000	0.0133	0.0124	0.0072	0.0210	0.4450
View p. Sub	0.0000	0.0000	0.0000	0.0000	0.7110	0.4220	0.4750

Table 14. P-values of Pearson correlation part 2

	(Dis)like p. View	Readability	Tot Views	Subscr.	Tot videos	Video Length	View p. Sub
Views	0.0000	0.1380	0.0000	0.0000	0.4290	0.0001	0.0000
Likes	0.0000	0.0001	0.0000	0.0000	0.0328	0.0000	0.0000
Dislike	0.0003	0.9680	0.0000	0.0000	0.2330	0.0133	0.0000
Comment	0.0000	0.0015	0.0000	0.0000	0.3160	0.0124	0.0000
Rating	0.0000	0.5500	0.2040	0.0002	0.0000	0.0072	0.7110
Title Len.	0.0000	0.0000	0.0000	0.0039	0.2280	0.0210	0.4220
Com p. View	0	0.6420	0.0000	0.0001	0.2410	0.4450	0.4750
(Dis)like p. View	1.0000	0.2350	0.0047	0.0000	0.0000	0.1670	0.4340
Readability	0.2350	1.0000	0.2610	0.2620	0.9610	0.0644	0.5510
Tot Views	0.0047	0.2610	1.0000	0	0.0000	0.0005	0.7230
Subscr.	0.0000	0.2620	0	1.0000	0.0000	0.0003	0.5800
Tot videos	0.0000	0.9610	0.0000	0.0000	1.0000	0.4410	0.6920
Video Length	0.1670	0.0644	0.0005	0.0003	0.4410	1.0000	0.4830
View p. Sub	0.4340	0.5510	0.7230	0.5800	0.6920	0.4830	1.0000

Table 15. Spearman correlation values part 1

	Views	Likes	Dislike	Comment	Rating	Title Len.	Com p. View
Views	1.0000	0.9040	0.9160	0.7490	-0.0119	-0.0129	-0.2980
Likes	0.9040	1.0000	0.9030	0.8340	0.2380	-0.0326	-0.0593
Dislike	0.9160	0.9030	1.0000	0.7800	-0.1490	0.0045	-0.1360
Comment	0.7490	0.8340	0.7800	1.0000	0.1490	-0.0203	0.3030
Rating	-0.0119	0.2380	-0.1490	0.1490	1.0000	-0.0831	0.1860
Title Len.	-0.0129	-0.0326	0.0045	-0.0203	-0.0831	1.0000	-0.0016
Com p. View	-0.2980	-0.0593	-0.1360	0.3030	0.1860	-0.0016	1.0000
(Dis)like p. View	-0.1490	0.2300	-0.0096	0.2390	0.5760	-0.0520	0.6230
Readability	0.0516	0.0470	0.0481	0.0574	0.0105	0.2370	0.0049
Tot Views	0.5940	0.5650	0.5560	0.4840	0.0102	0.0212	-0.1280
Subscr.	0.5720	0.6190	0.5670	0.5390	0.1130	0.0195	-0.0043
Tot videos	0.0111	-0.0074	0.0313	0.0484	-0.1060	0.1040	0.0562
Video Length	0.0033	0.0450	0.0917	0.1220	-0.1330	0.0881	0.2230
View p. Sub	0.3680	0.2710	0.3040	0.2050	-0.0475	-0.0745	-0.2400

Table 16. Spearman correlation values part 2

	(Dis)like p. View	Readability	Tot Views	Subscr.	Tot videos	Video Length	View p. Sub
Views	-0.1490	0.0516	0.5940	0.5720	0.0111	0.0033	0.3680
Likes	0.2300	0.0470	0.5650	0.6190	-0.0074	0.0450	0.2710
Dislike	-0.0096	0.0481	0.5560	0.5670	0.0313	0.0917	0.3040
Comment	0.2390	0.0574	0.4840	0.5390	0.0484	0.1220	0.2050
Rating	0.5760	0.0105	0.0102	0.1130	-0.1060	-0.1330	-0.0475
Title Len.	-0.0520	0.2370	0.0212	0.0195	0.1040	0.0881	-0.0745
Com p. View	0.6230	0.0049	-0.1280	-0.0043	0.0562	0.2230	-0.2400
(Dis)like p. View	1.0000	-0.0040	-0.0375	0.1490	-0.0320	0.1100	-0.2330
Readability	-0.0040	1.0000	0.0191	0.0158	-0.0037	-0.0105	0.0734
Tot Views	-0.0375	0.0191	1.0000	0.9070	0.6050	0.0092	-0.3140
Subscr.	0.1490	0.0158	0.9070	1.0000	0.4990	0.0768	-0.2490
Tot videos	-0.0320	-0.0037	0.6050	0.4990	1.0000	0.1050	-0.4970
Video Length	0.1100	-0.0105	0.0092	0.0768	0.1050	1.0000	-0.0866
View p. Sub	-0.2330	0.0734	-0.3140	-0.2490	-0.4970	-0.0866	1.0000

Table 17. P-values of Spearman correlation part 1

	Views	Likes	Dislike	Comment	Rating	Title Len.	Com p. View
Views	1.0000	0	0	0	0.2610	0.2230	0.0000
Likes	0	1.0000	0	0	0.0000	0.0021	0.0000
Dislike	0	0	1.0000	0	0.0000	0.6720	0.0000
Comment	0	0	0	1.0000	0.0000	0.0561	0.0000
Rating	0.2610	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
Title Len.	0.2230	0.0021	0.6720	0.0561	0.0000	1.0000	0.8810
Com p. View	0.0000	0.0000	0.0000	0.0000	0.0000	0.8810	1.0000
(Dis)like p. View	0.0000	0.0000	0.3680	0.0000	0	0.0000	0
Readability	0.0000	0.0000	0.0000	0.0000	0.3230	0.0000	0.6440
Tot Views	0	0	0	0	0.3370	0.0462	0.0000
Subscr.	0	0	0	0	0.0000	0.0655	0.6890
Tot videos	0.2940	0.4880	0.0032	0.0000	0.0000	0.0000	0.0000
Video Length	0.7590	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
View p. Sub	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 18. P-values of Spearman correlation part 2

	(Dis)like p. View	Readability	Tot Views	Subscr.	Tot videos	Video Length	View p. Sub
Views	0.0000	0.0000	0	0	0.2940	0.7590	0.0000
Likes	0.0000	0.0000	0	0	0.4880	0.0000	0.0000
Dislike	0.3680	0.0000	0	0	0.0032	0.0000	0.0000
Comment	0.0000	0.0000	0	0	0.0000	0.0000	0.0000
Rating	0	0.3230	0.3370	0.0000	0.0000	0.0000	0.0000
Title Len.	0.0000	0.0000	0.0462	0.0655	0.0000	0.0000	0.0000
Com p. View	0	0.6440	0.0000	0.6890	0.0000	0.0000	0.0000
(Dis)like p. View	1.0000	0.7050	0.0004	0.0000	0.0025	0.0000	0.0000
Readability	0.7050	1.0000	0.0724	0.1350	0.7270	0.3220	0.0000
Tot Views	0.0004	0.0724	1.0000	0	0	0.3890	0.0000
Subscr.	0.0000	0.1350	0	1.0000	0	0.0000	0.0000
Tot videos	0.0025	0.7270	0	0	1.0000	0.0000	0
Video Length	0.0000	0.3220	0.3890	0.0000	0.0000	1.0000	0.0000
View p. Sub	0.0000	0.0000	0.0000	0.0000	0	0.0000	1.0000