Explanatory Data Analysis to Evaluate Keyword Searches for Educational Videos on YouTube with a Machine Learning

Mambang[1)*], Ahmad Hidayat[2)], Johan Wahyudi[3)], Finki Dona Marleny[4)]
[1)2)]Universitas Sari Mulia, Banjarmasin, [3]STMIK Indonesia Banjarmasin, [4]Universitas Muhammadiyah Banjarmasin
[1)]mambang@unism.ac.id, [2)]ayat5621@gmail.com, [3)]johan77@stmik.id, [4)]finkidona@umbjm.ac.id

Submitted : July 1, 2022 | **Accepted** : July 19, 2022 | **Published** : July 19, 2022

Abstract: One of the most important parts of data science is the process of explanatory data analysis. This study aims to analyze learning videos on YouTube using search keywords such as learning biology, chemistry, physics, computers, mathematics, management, accounting, citizenship, history, and culture. The method used is the explanatory data analysis technique with a Machine Learning approach. The dataset used in this study uses learning video search keywords found on the YouTube digital platform. After doing a thorough analysis of all existing variables, we found that in the context of searching for learning video keywords on YouTube, the viewing variable has a heatmap correlation of 0.97 on the likes variable, 0.97 on the subscribers variable, -0.15 on the duration variable and 0.95 on the comment variable. The duration variable negatively correlates with all variables based on the analysis using a correlation heatmap using the seaborn library. Our analysis found that the number of learning videos with the search keyword Mathematics had the highest number of views among other variables. Further research can use existing variables or also add variables and add search keywords on YouTube. The data analysis approach can also be done using SPSS, R and also a Machine Learning approach with different libraries.

***Keywords:* Explanatory Data Analysis, Evaluating, Educational Videos, YouTube, Machine Learning**

INTRODUCTION

YouTube as an open digital platform provides many conveniences in its use. Content analysis and video quality can be learned to impart new knowledge to users (Foster et al. 2022). Data from *"We Are Social"* a company that focuses on reporting on the development of the internet, social media and technological developments, ranks Indonesia as the top three largest YouTube users in the world. The number of YouTube users in Indonesia is 127 million users. User involvement in accessing learning videos on Youtube provides data input on the number of views, likes, subscribers, duration and comments. YouTube digital platform provides easy access and as one of the supports for distance education (Al-zaman 2022), (Elareshi et al. 2022).

This study aims to analyze learning videos found on YouTube using search keywords such as learning biology, chemistry, physics, computers, mathematics, management, accounting, citizenship, history, and culture. From keyword searches carried out on learning videos on YouTube, analysis was carried out on all these keywords by analyzing the variables of the number of views, likes, subscribers, duration and comments to find correlations between all variables. Analysis was also performed on ten keywords used on searches on YouTube to find the most number of views, likes, subscribers, duration and comments on all keywords related to learning videos. Previous research discussing video analysis on YouTube as conducted by Negar Mohammadhassan, who analyzed and examined the effects and drivers in improving the quality of comments in watching videos on YouTube. The results obtained from this study state that video quality greatly affects engagement in commenting on videos on YouTube (Mohammadhassan, Mitrovic, & Neshatian 2022). In 2021, Atik Ramadhani conducted research on quality and reliability video halitosis on YouTube as a source of information. The results of this study stated that there was no difference in the number of video viewers at a video length of less than or more than four minutes (Ramadhani et al. 2021). Isil Yurdaisik in her research made an analysis of the first video seen on YouTube related to breast cancer. The results of this research found that of the 50 videos carried out breast

*name of corresponding author

cancer-related analysis, as many as 14% of videos were uploaded by doctors, 26% by health channels and 20% by patients and other videos were uploaded by various sources (Yurdaisik 2020). The process carried out in this study is by making an analysis of learning video keywords contained on YouTube. Data analysis and mining can be processed with a Machine Learning approach (Ahmed, Al-hamadani, & Satam 2022).

METHOD

This research uses explanatory data analysis techniques with a Machine Learning approach. Some of the libraries or supporting components in this analysis such as Anaconda navigator, Jupyter notebook, Pandas, Matplotlib, Seaborn and Numpy.

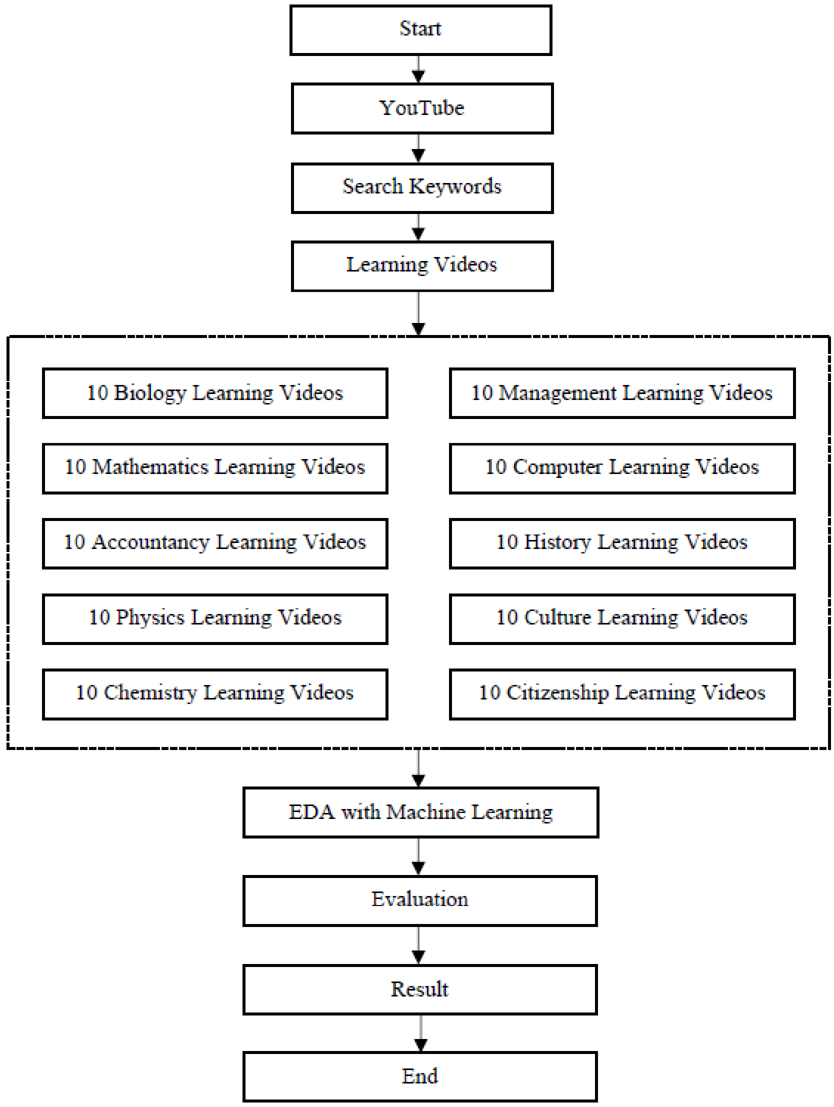
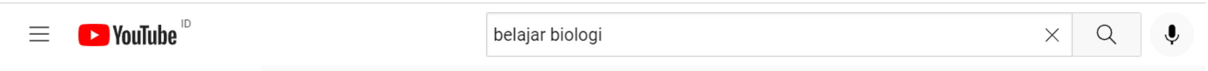

Start
YouTube
Search Keywords
Learning Videos
10 Biology Learning Videos
10 Management Learning Videos
10 Mathematics Learning Videos
10 Computer Learning Videos
10 Accountancy Learning Videos
10 History Learning Videos
10 Physics Learning Videos
10 Culture Learning Videos
10 Chemistry Learning Videos
10 Citizenship Learning Videos
EDA with Machine Learning
Evaluation
Result
End

Fig 1. Flowchart Research Methods

Figure 1 shows the stages of this study. There are ten search keywords that are carried out by the analysis process. Each of our learning video keywords collects ten videos on the first page of YouTube.



belajar biologi

Fig 2. Learning video keyword search

After the keywords for learning biology as in figure 2 are inputted in the YouTube search menu, as many as ten videos related to learning biology videos are used in this analysis process. All keywords are carried out the same process in the collection of videos that are the object of analysis.

*name of corresponding author

1. *Dataset collection*

The dataset (Elisawati et al.,2022) used in this study uses learning video search keywords found on the YouTube digital platform.

Table 1. Ten Keyword Channels Learning Biology Videos

Views	Like	Subscribers	Duration	Commentary
225000	3100	281000	4	88
15586	519	1870	110	15
760	47	383	4	16
31521	2100	1030000	9	49
1291	77	4790	4	6
1021000	31000	3710000	10	2348
13665	605	3290	7	32
101713	3400	125000	3	148
126416	3000	264000	7	84
270804	8800	176000	27	215

Table 1 shows a sample dataset from ten biology learning video channels used in this analysis. There are ten search keywords used to collect learning video datasets consisting of biology, chemistry, physics, computers, mathematics, management, accounting, citizenship, history, and culture. Each learning video keyword was collected by ten channels related to the learning video search keyword. Types of data are available with a variety of variables that can be used as new information (Zhao et al. 2021).

2. *Explanatory Data Analysis*

One of the most important parts of data science is the explanatory process of data analysis. Making analysis of the amount of data and recognizing the characteristics of data in data science also requires libraries contained in the python programming language (Muhammadiah, Wahab, and Surahman 2022). A series of initial data testing processes is a stage of explanatory data analysis that aims to identify patterns, anomaly data, hypothesis testing, and identify initial information. Predictive analytics by utilizing explanatory data with algorithms becomes very important in digging for information for new artificial intelligence (Davazdahemami, Zolbanin, & Delen 2022). The combination of relevant theories and analyses provides new information and knowledge (Zhao & You 2021). There are several Exploratory Data Analysis techniques that are often used such as Univariate Non-Graphical, Univariate Graphical, Multivariate Non-Graphical, Multivariate Graphical.

3. *Machine Learning*

Machine Leaning Approach (Minn 2022) in this method by using several libraries contained in python programming such as Pandas, Numpy, Matplotlib and also Seaborn. With Machine Learning, data can be developed with many models aimed at predicting data on many fields (Shi et al. 2022). Studying data with a computer is one of the goals of Machine Learning which is one of the branches of Artificial Intelligence.

RESULT

Explanatory data analysis results with Machine Learning were conducted on all search keywords consisting of the keywords learning biology, chemistry, physics, computer, mathematics, management, accounting, citizenship, history, and culture to evaluate the correlation or strong relationship of all variables consisting of the variables views, likes, subscribers, duration and comments.

Table 2. Keyword correlation results for learning video search using the Seaborn heatmap library

Keywords	Variable		Correlation Results
Biology	Like	Views	0.99
Chemistry	Like	Views	0.99
Physics	Commentary	Views	0.99
Computer	Like	Views	0.99

*name of corresponding author

Mathematics	Like	Commentary	0.84
Management	Like	Views	0.92
Accountancy	Like	Commentary	0.92
Citizenship	Like	Commentary	0.99
History	Like	Views	0.94
Culture	Like	Views	0.99

Variables from the keywords for searching for biology learning videos show that the view variable has a correlation with the like variable. The correlation value of the two variables is 0.99. The keyword chemistry learning video also has the same value of 0.99. The keyword physics learning video has a value of 0.99 with a different variable, namely comments with views. The keyword computer learning video has a value of 0.99. The keyword management learning video has a correlation value of 0.92. The keyword accounting learning video with the variable like with comments has a value of 0.92. Civic learning video keywords with a score of 0.99. The keyword history learning video with like and view variables has a value of 0.94. Cultural learning video keywords with a correlation value of 0.99. There are six learning video search keywords on YouTube that have the same variable between likes and views in correlation, namely in the keywords Biology, Chemistry, Computers, Management, History and Culture. There are three variables that are the same between likes and comments, namely in the keywords Mathematics, Accounting and Citizenship. There is one keyword with comment and display variables, namely on the physics keyword.

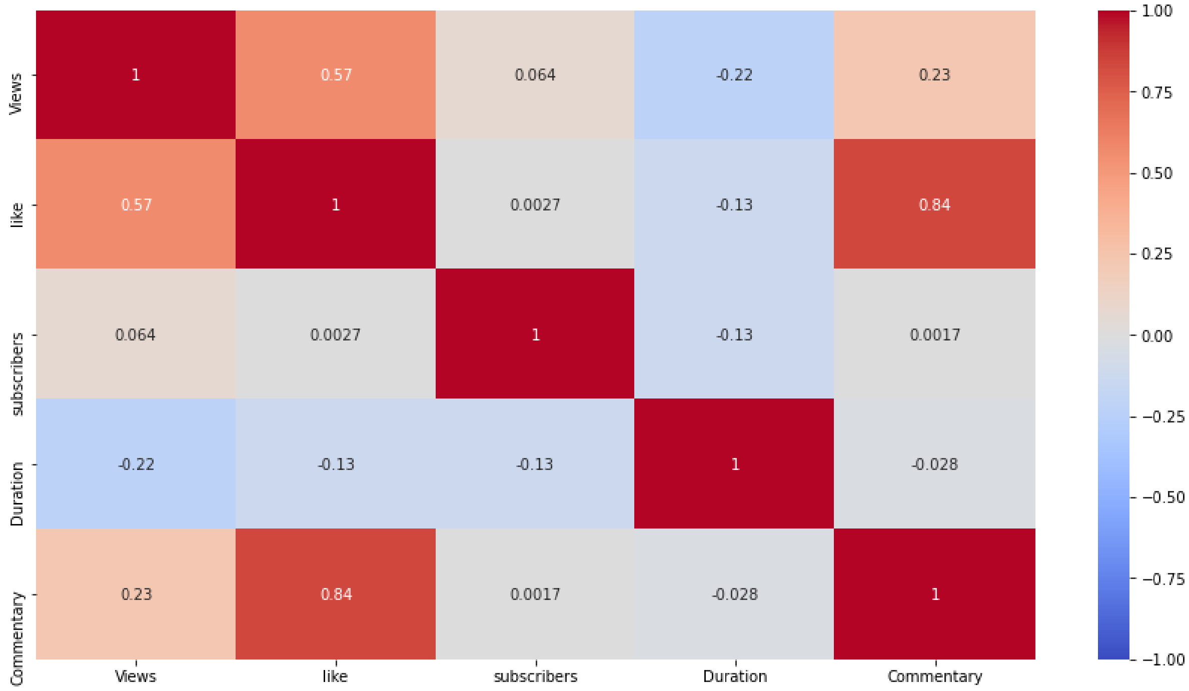


Fig 3. Keyword correlation heatmap of math learning search

Figure 3 shows the results of one of the correlation heatmaps on the keywords of searching for mathematics learning videos. Ten learning video keywords analyzed with a machine learning approach are carried out the same process as in figure 3.

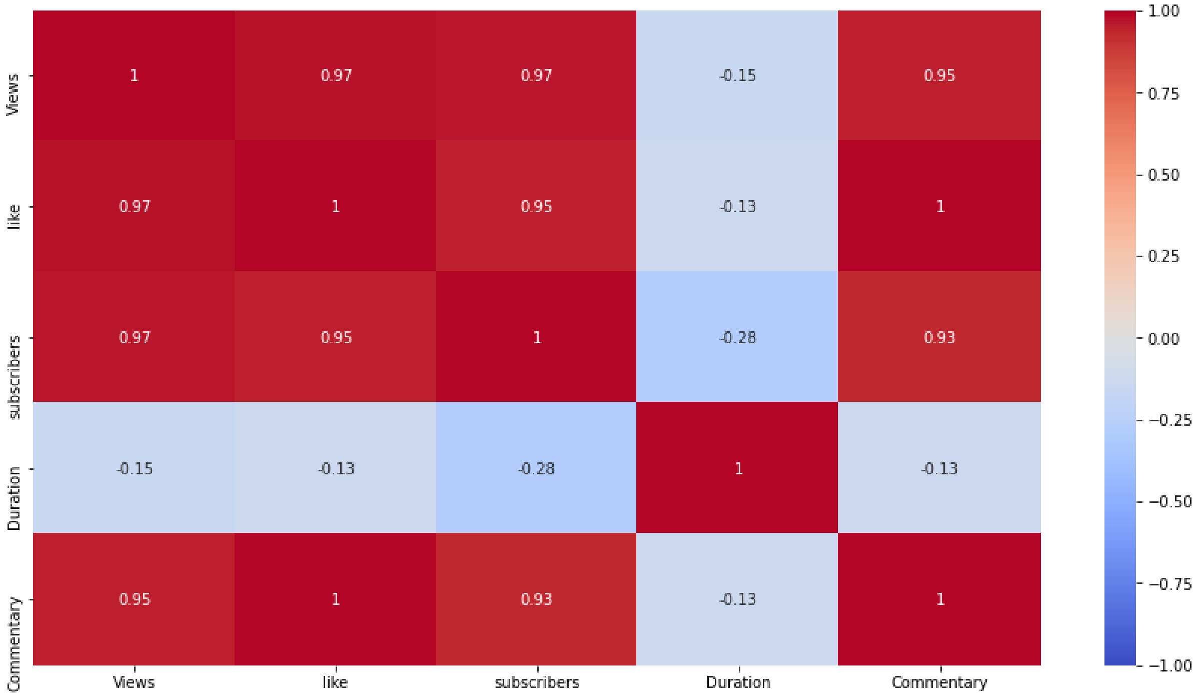


Fig 4. Heatmap correlation of ten keywords search learning video

Figure 4 shows the results of the correlation heatmap on ten keywords for learning video searches. Each of the learning video keywords is summed up on each variable such as the overall number of views, likes, subscribers, duration and also comments. After a thorough analysis of all existing variables, we found that in the context of searching for learning video keywords on YouTube, the view variable has a correlation heatmap of 0.97 on the like variable, 0.97 on the subscribers variable, -0.15 on the duration variable and 0.95 on the comment variable. Variables that have a correlation of 1.00 are found in the like with comment variables. Duration variables have a negative correlation to all variables based on analysis using the correlation heatmap using the seaborn library.

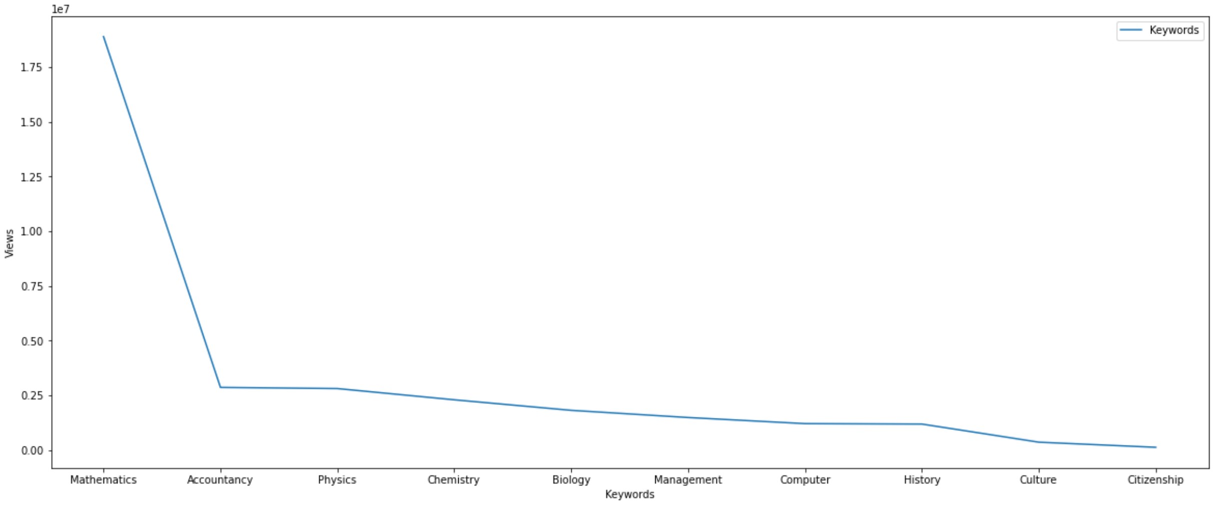


Fig 5. Lineplot keywords with the number of views.

Figure 5 shows the lineplot of the number of keywords with the number of views. Our analysis found that the number of learning videos with the keyword Math search had the most views among other variables. The result of combining ten channels related to the keyword Mathematics was at 18,901,838 views.

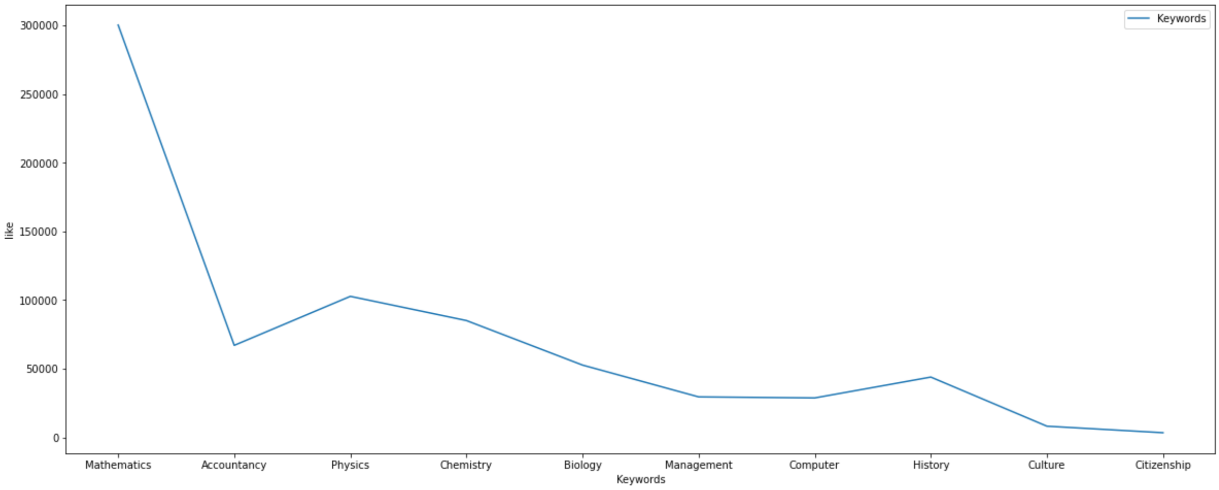


Fig 6. Lineplot keywords by number of likes

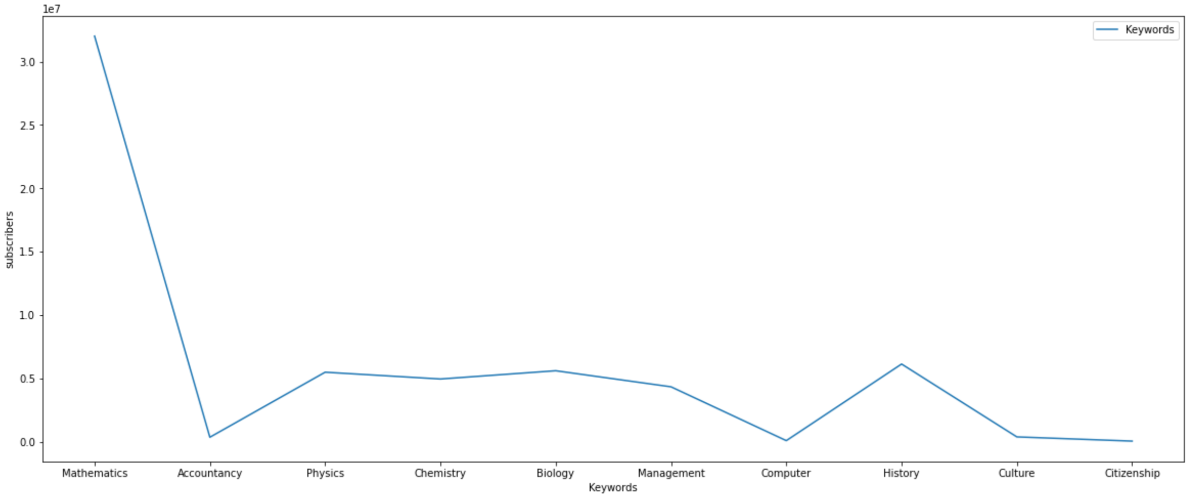


Fig 7. Lineplot keywords by the number of subscibers

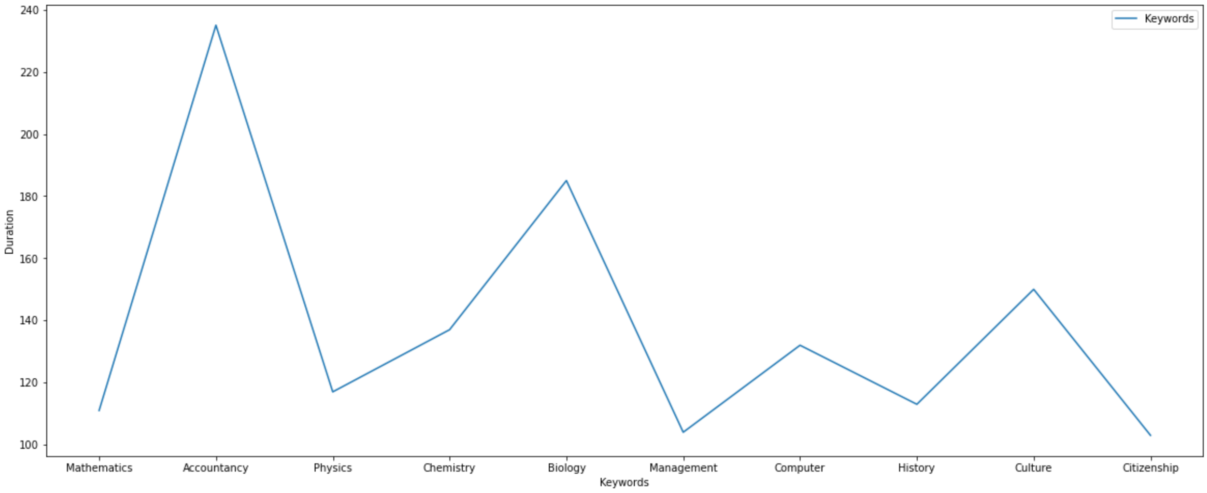


Fig 8. Lineplot keywords by number of durations

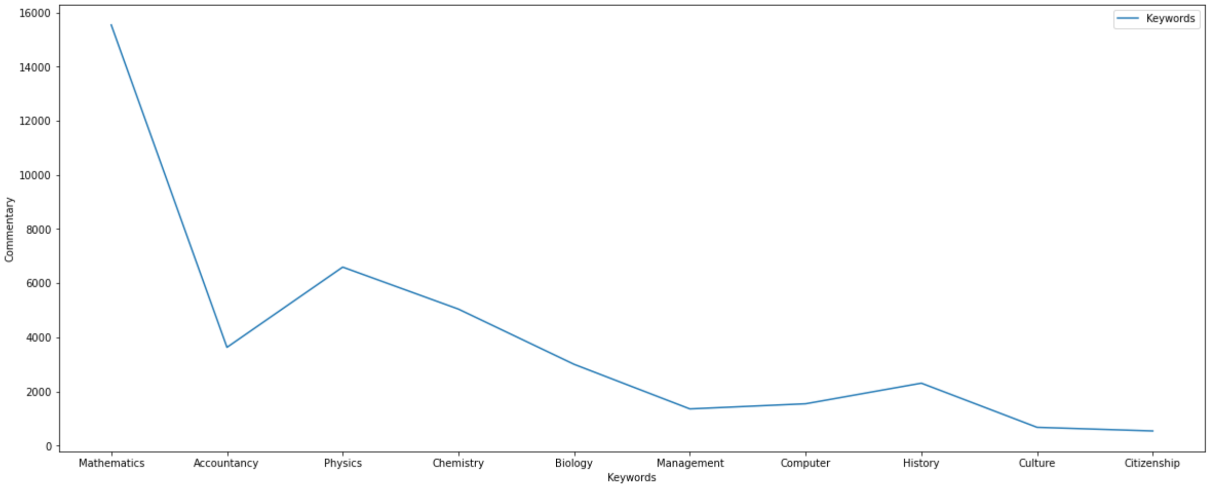


Fig 9. Lineplot keywords by number of comments

DISCUSSIONS

Explanatory data analysis with a Machine Learning approach aims to evaluate the keywords for searching learning videos on YouTube. A comprehensive evaluation of all variables showed that impressions correlated 0.97 with the like variable, subscribers did not correlate with the duration variable, the duration of time had a correlation in comments with a value of 0.95 and the like variable had a very strong correlation with comments. Furthermore, from the evaluation, it was found that the time duration variable in the learning video did not correlate with four other variables: impressions, likes, subscribers and comments. In making effective learning videos, there is no need to make the time duration of the learning video content an indicator.

Our research is also in line with Alyssa P. Lawson (Lawson and Mayer 2022), which aims to evaluate the effectiveness of the learning videos used when an instructor provides the material. Predictions on the popularity of YouTube videos can also be done with algorithms such as XGBoost (Nisa et al., 2021). Prediction of the number of views on YouTube videos can be done on all video content, be it educational videos, tourism, business and other fields (Kim and Lee 2021).

CONCLUSION

Ten keywords were used in this analysis with five variables Impressions, subscribers, likes, duration and comments. The likes with comments variable are strongly correlated with all learning video keywords. Keyword Math learning videos find the highest number of views, likes, subscribers, and comments. As for the variable duration, the highest number is found in the keywords of the Accounting learning video. Can add other learning video search variables and keywords for further research. We also advise subsequent research to analyze data using SPSS, R and Machine Learning approaches with different libraries for further research.

ACKNOWLEDGMENT

Thank you to LPPM Universitas Sari Mulia Banjarmasin for providing support both in the form of funding and also support for the provision of hardware to complete this research. Thank you also to colleagues at the University of Muhammadiyah Banjarmasin and STMIK Indonesia Banjarmasin for contributing to the completion of this research.

REFERENCES

Ahmed, Alaa H., Mokhaled N. A. Al-hamadani, and Ihab A. Satam. 2022. “Prediction of COVID-19 Disease Severity Using Machine Learning Techniques.” *Bulletin of Electrical Engineering and Informatics* 11(2):1069–74. doi: 10.11591/eei.v11i2.3272.

Al-zaman, Sayeed. 2022. “Social Mediatization of Religion : Islamic Videos on YouTube.” *Heliyon* 8(February):e09083. doi: 10.1016/j.heliyon.2022.e09083.

Davazdahemami, Behrooz, Hamed M. Zolbanin, and Dursun Delen. 2022. “An Explanatory Analytics Framework for Early Detection of Chronic Risk Factors in Pandemics.” *Healthcare Analytics* 2(January):100020. doi: 10.1016/j.health.2022.100020.

Elareshi, Mokhtar, Mohammed Habes, Enaam Youssef, Said A. Salloum, and Raghad Alfaisal. 2022. “SEM-ANN-Based Approach to Understanding Students ’ Academic-Performance Adoption of YouTube for Learning during Covid.” *Heliyon* 8(May 2021):e09236. doi: 10.1016/j.heliyon.2022.e09236.

*name of corresponding author

Elisawati, Linarta, Arie, Al Malikul, Ikhwanda Putra, and Herris Elvaningsih. 2022. "Analysis of Backpropagation Method in Predicting Drug Stock." *Sinkron : Jurnal Dan Penelitian Teknik Informatika* 7(2). doi: 10.33395/sinkron.v7i2.11269.

Foster, Brian K., William Mack Malarkey, Timothy C. Maurer, Daniela F. Barreto Rocha, Idorenyin F. Udoeyo, and Louis C. Grandizio. 2022. "Biceps Tendon Rupture Videos on YouTube : An Analysis of Video Content and Quality." *Journal of Hand Surgery Global Online* 4(1):3–7. doi: 10.1016/j.jhsg.2021.10.009.

Minn, Sein. 2022. "AI-Assisted Knowledge Assessment Techniques for Adaptive Learning Environments." *Computers and Education: Artificial Intelligence* 3(July 2021):100050.

Mohammadhassan, Negar, Antonija Mitrovic, and Kourosh Neshatian. 2022. "Investigating the Effect of Nudges for Improving Comment Quality in Active Video Watching." *Computers & Education* 176:104340. doi: 10.1016/j.compedu.2021.104340.

Muhammadiah, Mas, Abdul Wahab, and Susilo Surahman. 2022. "Development of Web-Based Learning Evaluation Tools in Vocational High Schools." *Sinkron : Jurnal Dan Penelitian Teknik Informatika* 7(2):308–13. doi: 10.33395/sinkron.v7i2.11292.

Ramadhani, Atik, Zenobia Zettira, Yuanita Lely Rachmawati, and Ninuk Hariyani. 2021. "Quality and Reliability of Halitosis Videos on YouTube as a Source of Information." *Dentistry Journal Article* 9(10):1–9. doi: 10.3390/dj9100120.

Shi, Hui, Dong Yang, Kaichen Tang, Chunmei Hu, Lijuan Li, and Linfang Zhang. 2022. "Explainable Machine Learning Model for Predicting the Occurrence of Postoperative Malnutrition in Children with Congenital Heart Disease." *Clinical Nutrition* 41(1):202–10. doi: 10.1016/j.clnu.2021.11.006.

Yurdaisik, Isil. 2020. "Analysis of the Most Viewed First 50 Videos on YouTube about Breast Cancer." *BioMed Research International* 2020:1–7. doi: 10.1155/2020/2750148.

Zhao, Pengxiang, He Haitao, Aoyong Li, and Ali Mansourian. 2021. "Impact of Data Processing on Deriving Micro-Mobility Patterns from Vehicle Availability Data." *Transportation Research Part D* 97(June):102913. doi: 10.1016/j.trd.2021.102913.

Zhao, Yanmin, and Yang You. 2021. "Design and Data Analysis of Wearable Sports Posture Measurement System Based on Internet of Things." *Alexandria Engineering Journal* 60(1):691–701. doi: 10.1016/j.aej.2020.10.001.

*name of corresponding author