YouTube Trend Analysis

Arushi Pathik
Computer Engineering
MPSTME, SVKM's NMIMS
Mumbai,India
pathikarushi2000@gmail.com

Saumya Patni
Computer Engineering
MPSTME, SVKM's NMIMS
Mumbai,India
saum.patni@gmail.com

Vaibhav Patel
Computer Engineering
MPSTME, SVKM's NMIMS
Mumbai,India
vaibhav6644@gmail.com

Jash Patel
Computer Engineering
MPSTME, SVKM's NMIMS
Mumbai,India
jashpatel2000@gmail.com

Artika Singh
Computer Engineering
MPSTME, SVKM's NMIMS
Mumbai,India
artika.singh@nmims.edu.in

***Abstract*—Nowadays, Online video streaming services are extremely popular. YouTube give facility to their content creators to spread their knowledge, thoughts, and interesting content with users. In YouTube there is a trending section which shows currently most popular videos, ensuring that a video reaches the widest possible audience. Other than those videos rest are unpredictable, with the exception of few viral videos having a large number of views and are guaranteed to be in the trending section. Data analysis and Data mining are critical in today's world, and businesses are improving their operations by using social media. The aim of paper is to investigate YouTube's trending videos data. Users in the app use Views, Comments, Likes, and Dislikes. Classification algorithms like Linear Regression, Decision Tree, many other Machine Learning models can be used by using Python libraries like pandas and matplotlib, to classify and analyze YouTube data, as well as collect useful information.**

***Index Terms*—YouTube Trend, Machine Learning, Python, Linear Regression, Social Media, Views**

I. Introduction

In the past decade, we've witnessed major advances in the technological field of the world. One such revolutionizing discovery was YouTube. An online video streaming platform, YouTube, has managed to grow and expand into the world to become the most used online video platform today. A section called "trending" is especially added in YouTube which is getting updated continuously.

By analyzing trending videos content creators can get new insights and ideas to make their channels grow in terms of popularity and makes as brand. Businesses and companies that uses digital platforms and social media can get benefits from this, by posting videos on them or sponsoring relevant channels at the right time through analysis. [6]

This paper aids in analysing parameters that result in a certain video content being tagged as 'viral'.

1.1 Objectives

- To understand what parameters aid in placing a video in YouTube's trending section and to gain an understanding of YouTube's working algorithm. Also to identify crucial components that can be leveraged to help YouTube channels develop.
- Provide suggestions to content creators organizations to expand their video's audience by explaining what parameters can be improved upon to get their video trending.
- Develop a detailed result analysis by evaluating different algorithms and analyze their accuracy, prediction strength, and consistency.
- To help future content creators by providing a platform to gauge their viral quotient before experimenting with YouTube's platform.

II. Methods of analyzing

Analysis is done using 3 different techniques.

2.1 Evaluation of a pre-defined dataset of Trending and Non-Trending videos

The parameters of a pre-defined dataset can be used to help predict the viral quotient of a video. Views, likes, dislikes, comments, and, most importantly, the date and time of uploading are all factors. These variables will help in determining the fundamental requirements of a video to being in YouTube's trending section. Such kind of data can be fetched through YouTube's API. The API has a number of features, including the ability to download video files, descriptions, titles, thumbnails, and other data.

2.2 Difference between Viral and Non-Viral or Trending and Non-Trending videos

Parameters comparison of viral and non-viral videos is an important deciding factor while making further analysis. Deciding factors such as the threshold of the like numbers, views above which a video is tagged as 'viral', is used to segregate trending from non-trending. A good amount of data, in appropriate proportions, about viral and non-viral videos will make the better conclusion with more accurate results. [10]

2.3 Visual result analysis of the implemented algorithms

Three major Machine Learning and Data Mining algorithms are used: Decision Tree Algorithm, Logistic Regression, and Bayes Classifier (Gaussian). Further implementations of these algorithms on the dataset are evaluated by a detailed result analysis for visualization purposes. The visualization process aids a layman to understand the motive of the paper and grasping concrete knowledge about each algorithm's accuracy in real-time. Another benefit of the result visualization is that it helps to compare and draw conclusions while carefully analyzing each algorithm simultaneously.

We can develop much more concrete conclusions by combining the outcomes of these three methods of analysis, as each of these methods has its own set of advantages and disadvantages.

III. TRENDING NON-TRENDING VIDEOS: A CONCEPT

The YouTube Trending tab displays what's hot in their region, as well as which events or videos are receiving the most views. In the Trending section, videos with a large number of views and users who find the video suits to their interest are displayed. Some most popular videos, such as upcoming movie trailers or a new release song by a popular performer, are expected. Other videos are unexpected and are dependent on the video's time, event, and day. All users in each country see the same list of Trending videos; no customized videos are provided. Every 15 minutes, the list of videos in the Trending section is updated. Users cannot pay to have their videos appear in trending results, therefore trending videos are never paid advertisements.

On the other hand, non-trending videos are the ones that can have a greater number of dislikes, fewer views, comments, or even likes. These are the type of videos that people do not wish to see. However, not all non-trending videos are bad. Some may just be uploaded at the wrong time of the day, or which the wrong description, and hence missed its chance to fit in YouTube's algorithm.

IV. DATASET SELECTION & DATA CLEANING

In this paper, the entire analysis is done based on a dataset chosen by us. The dataset used for this paper's analysis was a total of 4500 video entries. The chosen dataset (uncleaned) was then loaded on and cleaned using various pre-processing techniques.

We used a dataset that we found on the internet. We'll look at both trending and non-trending videos from all across the world. After the entire analysis result declaration, we will implement the proposed model on a fresh set of videos, which are not termed as 'trending' and 'non-trending'. This is done to ensure the working of the proposed model and evaluate its accuracy.

Any study requires very clean and reliable data that is free of redundancies and missing values, therefore data cleaning is a critical phase in the process.
The Python libraries that we used for cleaning and formatting the data are pandas and NumPy.

We performed the following steps in the data cleaning process with pandas:

- Checking NULL values and deleting them
- Dropping unwanted columns
- Creating our Target Variable column

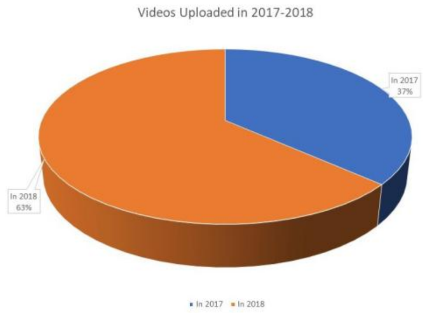


Fig. 1: Analysis of publishing videos in 2017-2018

V. ALGORITHMS

We have deeply used Machine Learning (ML) algorithms to develop our model and conclude our findings.

5.1 Decision Tree Algorithm

Decision Tree is a classification algorithm that can be used to perform many machine learning tasks like regression, classification, prediction, and multi-output operations. The Tree approach is a supervised learning model which are trained for resolving regression and classification problems. They are extremely powerful algorithms that can fit large datasets. A selection tree is frequently used to construct a trained model that predicts the target parameter's value or class by learning basic decision making rules from past training data. Decision Tree divide instances into two or more homogeneous sets based on the most informative splitter / separator in the input (independent) variables. This tree-structured classifier's intermediate nodes have record properties, branches show decision rules, and each leaf node gives a class (conclusion).

5.2 Logistic Regression

Logistic regression is used for predicting the discrete outcome probability of given an input (independent) variable. The most common logistic regression model yields a binary response, as shown below: B. Yes or no, true or false, and so on. When classification is difficult, logistic regression is a useful analytical method for determining whether a new sample properly fits the category. This enables newly submitted videos to be labelled as trending or non-trending.

5.3 Gaussian Naive Bayes Classifier

Gaussian Naive Bayes models all continuously scored features as a Gaussian (normal) distribution. Assume the data has a normal distribution with no covariance (independent dimensions) between the parameters in order to build a simple model. Simply compute the quality deviation and mean of the points in each label to fit this model. This is frequently all that is required to carry out this type of distribution.

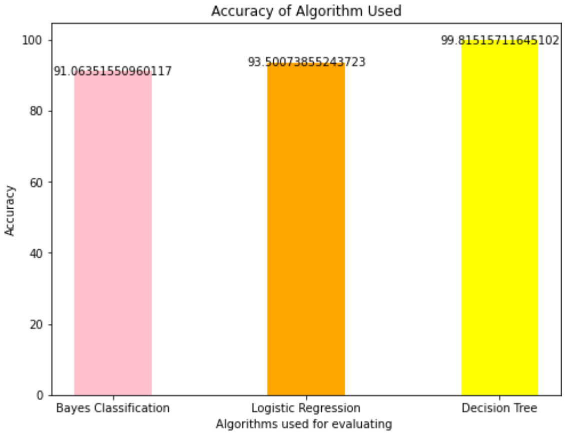


Fig. 2: Comparison: Accuracy of all algorithms

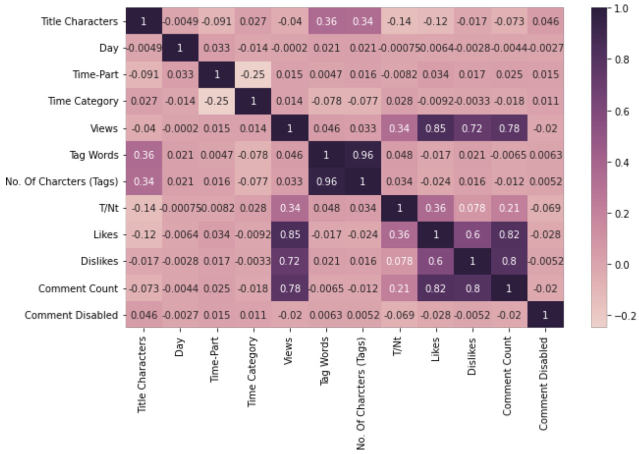


Fig. 3: Co-Relation Matrix

VI. Likes Count Analysis

A video can be marked as ‘trending’ based on the likes on it. It is natural to assume that a very high number of likes will automatically make a video good and hence trend it. Similarly, a high number of dislikes on the video will eventually not be liked by a large crowd and hence be tagged as ‘non-trending’. Hence, by marking a threshold for the number of likes, we can segregate trending and non-trending videos for further analysis.

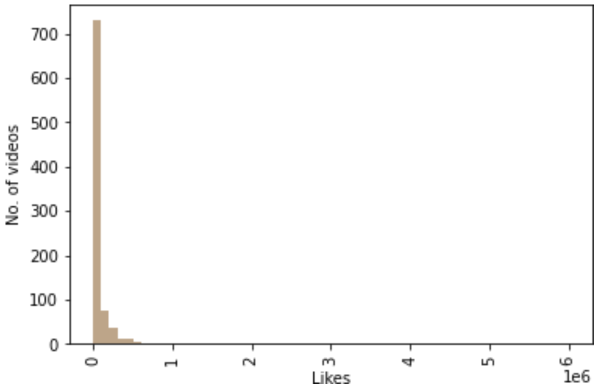


Fig. 4: Shows the number of likes for trending videos

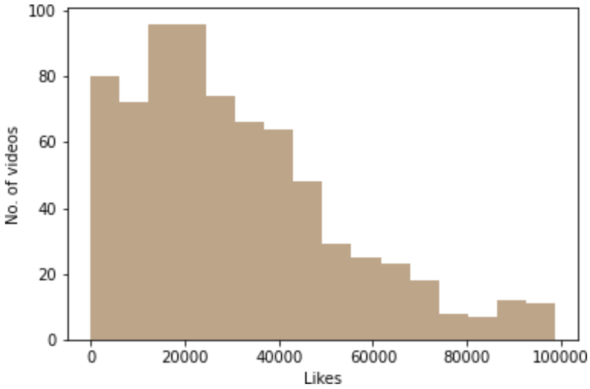


Fig. 5: Detailed likes distribution analysis

VII. View Count Analysis

Similar to the likes count, a trending video will eventually be watched by a large crowd and hence, will have a high number of views. In the chosen dataset, we have a wide range of the number of views in each video, therefore our dataset will give us a distinct result.

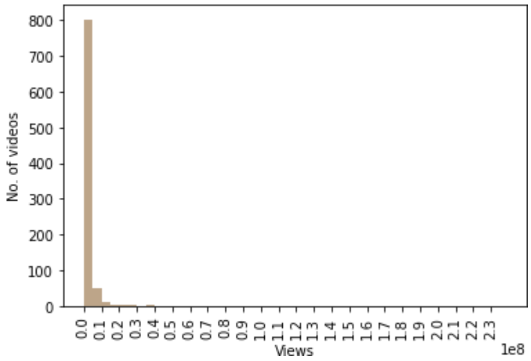


Fig. 6: Shows the views for trending videos

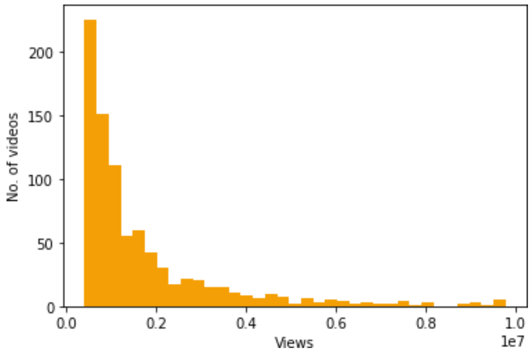


Fig. 7: Detailed views distribution analysis

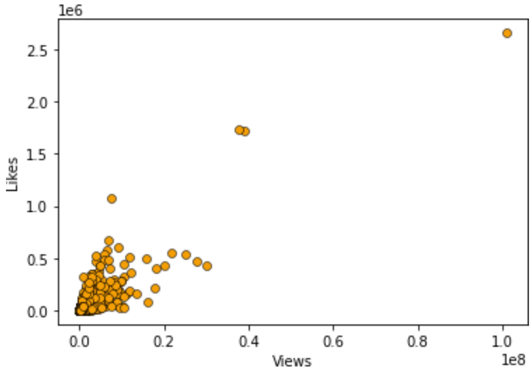


Fig. 8: Shows the relation between Likes & Views

VIII. Time of upload

By integrating the publishing time for each video with the amount of videos for each time slot, we can determine the best time for a trending video to be published and distributed. As a result, we can forecast whether a video broadcast in a specific time slot with a specific set of attributes will trend or not using algorithms like regression. In this paper, the entire day if divided into slots and assigned digits by range (morning, afternoon, evening, and night) for an enhanced understanding.

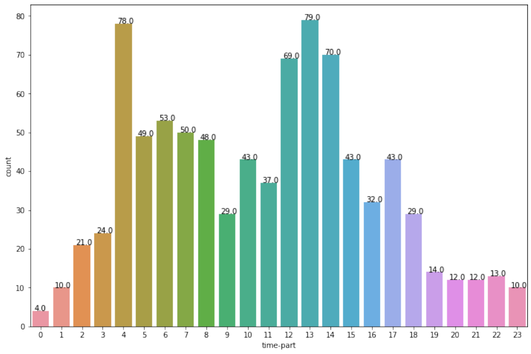


Fig. 9: Number of videos published per hour

IX. Day of upload

With a similar ideology as time of upload, the seven week days were assigned digits (Sunday being 0, Monday being 1, and so on). This was done to get a clear insight on which days are favourable to upload a video. This, combined with the perfect days' time slot, can ensure a high viral quotient.

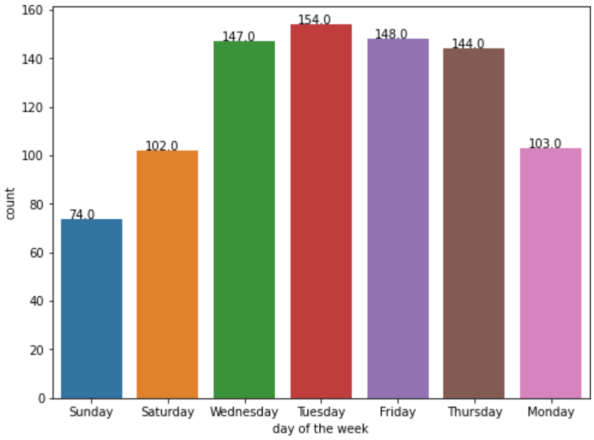


Fig. 10: Publishing Day

X. Video Title Length and Video Tags Analysis

We can retrieve the most common words in video titles using Python and the Pandas package after removing the normal structure words like articles, prepositions, and joining words such as the, a, an, but etc. Ignoring words like "the" and "of", we observe that the symbols "—" and "-" occurred several times in trending video titles. [6]

Hence, the number of words used in a video's title and tags used is Video description, is taken and considered for analysis.

- All trending videos have at least 10 characters in their Title.
- Highest Number of trending videos were found which had characters between 90-100.
- On average, trending videos have 21 tags and huge percentage of Trending Videos have 30 to 70 tags.
- Only 3.5 percent of trending videos have no tags.

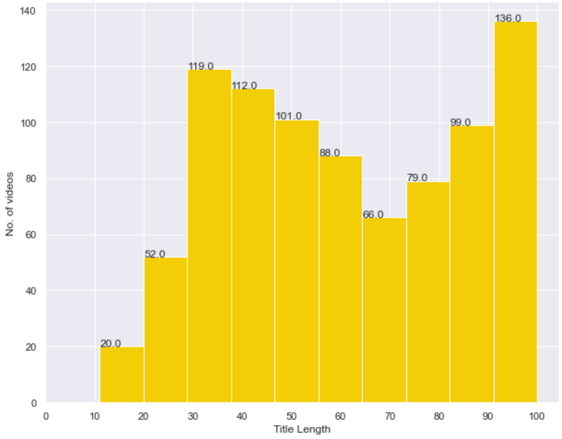


Fig. 11: Title character length for each count of trending videos

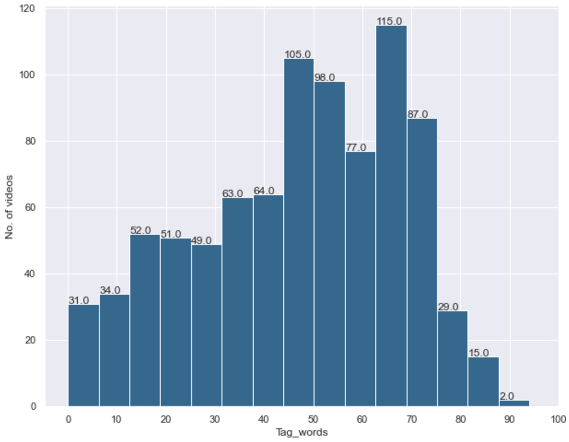
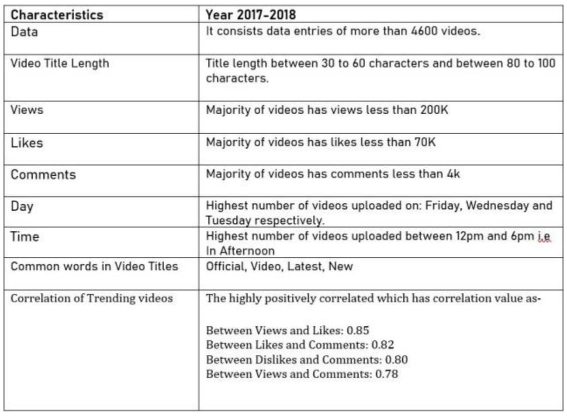


Fig. 12: Distribution Plot of Number of Tags

XI. Interpretation of some key results of Analysis.



Characteristics	Year 2017-2018
Data	It consists data entries of more than 4600 videos.
Video Title Length	Title length between 30 to 60 characters and between 80 to 100 characters.
Views	Majority of videos has views less than 200K
Likes	Majority of videos has likes less than 70K
Comments	Majority of videos has comments less than 4k
Day	Highest number of videos uploaded on: Friday, Wednesday and Tuesday respectively.
Time	Highest number of videos uploaded between 12pm and 6pm i.e In Afternoon
Common words in Video Titles	Official, Video, Latest, New
Correlation of Trending videos	The highly positively correlated which has correlation value as- Between Views and Likes: 0.85 Between Likes and Comments: 0.82 Between Dislikes and Comments: 0.80 Between Views and Comments: 0.78

Fig. 13: Characteristics Table

XII. Result Analysis

After doing Analysis on various Parameters of Trending Video, we have found out that:

- The Videos which have Title characters between 30-50 and 90-10 are more likely to Trend, as 45 percent of the Trending Videos have these stats.
- The Videos which have between 50 to 70 hashtags, are more Likely to Trend as 37 percent of the trending Videos have tags between 5-70.
- Tuesday is a favourable day of uploading a Video as it has more chances to Trend according to our data.
- The videos uploaded between 10 to 3pm are more likely to trend according to our Analysis.

Based on the above Analysis, we have created an Application which takes **Title length, number of tags, day and time of Video upload** as an input and gives the probability of Video going to Trending Page with respect to the highest chances possible and also gives appropriate feedback to the User according to the inputs with the help of which, he/she can do required changes and increase his chances of making the Video Trend.

We created this application with an Aim to help anyone who wants his video to Trend, and with the help of it, he/she can take care of the the above mentioned parameters which play a very important part of making the video trend.

XIII. Conclusion

Our findings for measuring, assessing, and comparing essential characteristics of YouTube popular videos were reported in this study. Knowing the optimal time to publish a video to YouTube isn't enough to get millions of views and make your video popular. Other elements to consider are good titles, thumbnails, video SEO, tagging, and the subscribers count, which are all important in increasing views for your material. Understanding these statistics will aid YouTube in not only developing better video processing algorithms, but also in making judgments for individual youtubers.

Therefore, in this paper, we have presented a theory to analyse and help predict the virality quotient of a certain video content. After detailed analysis, we have calculated the probability of the videos that have the potential to 'trend'. Assuming the highest chances of going viral to be 1, we have presented our findings.
Additionally, proper feedback is given to the users if the calculated probability is less than 1 to enhance their video's virality qoutient.

References

[1] A. Deza and D. Parikh, “Understanding image virality,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12-June-2015, pp. 1818–1826, 2015, doi: 10.1109/CVPR.2015.7298791.

[2] L. T. Rui, Z. A. Afif, R. D. R. Saedudin, A. Mustapha, and N. Razali, “A regression approach for prediction of Youtube views,” Bull. Electr. Eng. Informatics, vol. 8, no. 4, pp. 1502–1506, 2019, doi: 10.11591/eei.v8i4.1630.

[3] Q. Wang, F. Miao, G. K. Tayi, and E. Xie, “What makes online content viral? The contingent effects of hub users versus non–hub users on social media platforms,” J. Acad. Mark. Sci., vol. 47, no. 6, pp. 1005–1026, 2019, doi: 10.1007/s11747-019-00678-2.

[4] I. Mohr, “Going Viral: An Analysis of YouTube Videos,” J. Mark. Dev. Compet., vol. 8, no. 3, pp. 43–48, 2014, [Online]. Available: http://eds.b.ebscohost.com.libproxy.txstate.edu/eds/pdfviewer/pdfviewer?sid=4b676789-c238-4150-b0c8-204cc8c1e537

[5] J. Arthurs, S. Drakopoulou, and A. Gandini, “Researching YouTube,” Convergence, vol. 24, no. 1, pp. 3–15, 2018, doi: 10.1177/1354856517737222.

[6] S. Gayakwad, R. Patankar, and D. Mane, “Analysis on YouTube Trending Videos,” Int. Res. J. Eng. Technol., pp. 4247–4253, 2020, [Online]. Available: www.irjet.net.

[7] M. Fyfield, M. Henderson, and M. Phillips, “Navigating four billion videos: teacher search strategies and the YouTube algorithm,” Learn. Media Technol., vol. 46, no. 1, pp. 47–59, 2021, doi: 10.1080/17439884.2020.1781890.

[8] F. Figueiredo, J. M. Almeida, M. A. Goncalves, and F. Benevenuto, “On the dynamics of social media popularity: A you tube case study,” ACM Trans. Internet Technol., vol. 14, no. 4, 2014, doi: 10.1145/2665065.

[9] L. Jiang, Y. Miao, Y. Yang, Z. Lan, and A. G. Hauptmann, “Viral Video Style,” pp. 193–200, 2014, doi: 10.1145/2578726.2578754.

[10] I. Barjasteh, Y. Liu, and H. Radha, “Trending Videos: Measurement and Analysis,” no. January 2015, 2014, [Online]. Available: http://arxiv.org/abs/1409.7733.

[11] R. Amanda and E. S. Negara, “Analysis and Implementation Machine Learning for YouTube Data Classification by Comparing the Performance of Classification Algorithms,” J. Online Inform., vol. 5, no. 1, pp. 61–72, 2020, doi: 10.15575/join.v5i1.505.

[12] M. Airoldi, D. Beraldo, and A. Gandini, “Follow the algorithm: An exploratory investigation of music on YouTube,” Poetics, vol. 57, no. 2015, pp. 1–13, 2016, doi: 10.1016/j.poetic.2016.05.001.

[13] G. F. Khan and S. Vong, “Virality over youtube: An empirical analysis,” Internet Res., vol. 24, no. 5, pp. 629–647, 2014, doi: 10.1108/IntR-05-2013-0085.

[14] D. Han, “Predicting Influencer Virality on Twitter,” no. June, 2020.

[15] R. Andryani, E. S. Negara, and D. Triadi, “Social Media Analytics: Data Utilization of Social Media for Research,” J. Inf. Syst. Informatics, vol. 1, no. 2, pp. 193–205, 2019, doi: 10.33557/journalisi.v1i2.23.

[16] K. R. Purba, D. Asirvatham, and R. K. Murugesan, “An analysis and prediction model of outsiders percentage as a new popularity metric on Instagram,” ICT Express, vol. 6, no. 3, pp. 243–248, 2020, doi: 10.1016/j.icte.2020.07.001.

[17] B. Batrinca and P. C. Treleaven, “Social media analytics: a survey of techniques, tools and platforms,” AI Soc., vol. 30, no. 1, pp. 89–116, 2015, doi: 10.1007/s00146-014-0549-4.

[18] L. Jiang, Y. Miao, Y. Yang, Z. Lan, and A. G. Hauptmann, “Viral Video Style,” pp. 193–200, 2014, doi: 10.1145/2578726.2578754.

[19] M. Ba¨rtl, “YouTube channels, uploads and views: A statistical analysis of the past 10 years,” Convergence, vol. 24, no. 1, pp. 16–32, 2018, doi: 10.1177/1354856517736979.

[20] J. Arthurs, S. Drakopoulou, and A. Gandini, “Researching YouTube,” Convergence, vol. 24, no. 1, pp. 3–15, 2018, doi: 10.1177/1354856517737222.

[21] X. Cheng, C. Dale, and J. Liu, “Statistics and social network of YouTube videos,” IEEE Int. Work. Qual. Serv. IWQoS, pp. 229–238, 2008, doi: 10.1109/IWQOS.2008.32.

[22] R. Zhou, S. Khemmarat, and L. Gao, “The impact of YouTube recommendation system on video views,” Proc. ACM SIGCOMM Internet Meas. Conf. IMC, pp. 404–410, 2010, doi: 10.1145/1879141.1879193.