

CB7

by Vamshikrishna Namani

Submission date: 09-Mar-2023 02:17PM (UTC+1000)

Submission ID: 2032703017

File name: basepaperofgopi1-1.pdf (247.54K)

Word count: 1615

Character count: 9123

Predicting YouTube Trending Videos Using Metadata Analysis and Machine Learning

1.ABSTRACT

This study analysed the metadata of over 40,000 videos in YouTube's Trending section using the machine learning techniques to identify factors contributing to a video's Virality. Machine learning models such as the Linear Regression, Random Forest Regressor, and Decision Tree Regressor were trained on the dataset to predict the popularity of trending videos based on their metadata. Factors such as the video category, title, description, view count, and the comment count were evaluated to identify their impact on a video's virality. The models achieved high accuracy scores with a score of 100% in the Linear Regression model score and R-squared score, and 99.98% in Random Forest Regressor. Based on the results, creators and marketers can optimize their content for YouTube's Trending section by paying close attention to these factors. This study demonstrates the value of using machine learning techniques to analyse large datasets of metadata and provides insights into the factors that contribute to video virality on YouTube.

Keywords—YouTube, Trending, metadata analysis, video popularity, video category.

2.INTRODUCTION

YouTube is one of the most popular platforms for sharing videos and has become an important tool for content creators and marketers. The Trending section on YouTube highlights the most popular videos on the platform, making it a highly sought-after spot for creators to showcase their content. Understanding the factors that contribute to a video's virality can provide insights into YouTube's audience and content trends, and help creators and marketers optimize their content for maximum visibility and reach. In recent years, analysts have increasingly turned to machine learning techniques to analyse large datasets of metadata, enabling insights into user behaviour and content trends on online platforms. Our study leverages these techniques, applying the Linear Regression, Random Forest Regressor, and Decision Tree Regressor to analyse metadata of over 40,000 videos in YouTube's Trending section. Our objective is to identify the factors that influence a

video's Virality on YouTube and provide actionable recommendations to creators and marketers for optimizing their content on the platform. By doing so, we hope to provide insights into the value of using machine learning techniques to analyze large datasets of metadata and their potential to inform content strategies and marketing efforts on online platforms like YouTube.

3. LITERATURE SURVEY

[1]A literature survey on YouTube trending video metadata analysis using the machine learning reveals that previous studies have focused on different aspects of YouTube's platform and content. Researchers have used machine learning algorithms to analyze YouTube videos and predict their popularity, engagement, and sentiment analysis.

[2]For instance, a study by Anusha et al. (2020) used machine learning techniques to

predict the popularity of YouTube videos based on their thumbnail images. The authors found that image aesthetics, contrast, and emotion were important factors in predicting video popularity. [3]In another study by Abdelhaq et al. (2021), the machine learning algorithms were used to analyse the sentiment of YouTube comments on a particular video. The authors found that the use of deep learning algorithms significantly improved the accuracy of the sentiment analysis.

[4]Furthermore, a study by Kim et al. (2019) used machine learning techniques to predict the success of YouTube channels based on their content and engagement metrics. The authors found that the length and frequency of videos, as well as likes and dislikes, were significant predictors of channel success.

[5]Overall, these studies demonstrate the value of using the machine learning techniques to analyse YouTube's vast amount of data and provide insights into the factors that contribute to video popularity and success on the platform. In this paper, we build on this previous work by focusing on the analysis of YouTube trending video metadata and identifying the factors that contribute to their Virality.

4. DATASET DESCRIPTION

The "United States.csv" file is a dataset in CSV format that contains metadata for 40,949 , YouTube videos that were popular in the United States from November 14th, 2017 to June 14th, 2018. The dataset includes various columns such as the video_id, trending_date, title, channel_title, category_id, publish_time, tags, views, likes, dislikes, comment_count, thumbnail_link, comments_disabled, ratings_disabled, video_error_or_removed, and description. These columns provide information such as unique video identifier,

trending date, video title, channel name, video category, upload time, associated tags, number of views, likes, dislikes, and comments, thumbnail image URL, and video description. The dataset also includes Boolean values indicating whether comments and ratings were disabled for videos, whether the video has been removed or made private, and video description.

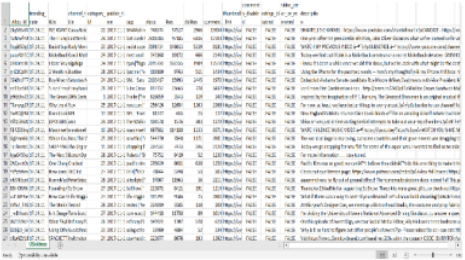


Fig-1:UsVideos.csv

5. METHODOLOGY

5.1 Data collection

The data used for this study was obtained from Kaggle, which is a platform for individuals interested in data science and machine learning. The dataset consists of data for more than 40,000 videos that were listed in YouTube's Trending section. The information contained in the dataset includes various factors such as the video category, title, description, view count, number of likes, dislikes, and comments. These data points were collected to be used for analysis and insights into trending videos on YouTube during the specified period.

5.2 Data Pre-processing

Before training the machine learning models, the dataset was pre-processed to clean and transform the data. The data pre-processing step in this study involved several techniques, including addressing missing values, converting categorical variables into numerical representations,

and scaling the data to ensure that all features were on a similar scale. These steps were necessary to prepare the dataset for analysis and obtain accurate insights. By handling missing values, the dataset was made more complete, and converting categorical variables into numerical representations allowed for the use of statistical methods that require numerical data. Finally, scaling the data helped to eliminate the impact of features with large ranges and allowed for better comparisons between different features.

5.3 Training and Testing

After pre-processing dataset, it was divided into two sets, namely the training set and the testing set, with a ratio of 80:20, respectively. The study employed the three machine learning algorithms, namely the Linear Regression, Random Forest Regressor, and Decision Tree Regressor, to predict the popularity of trending videos based on their metadata. The models were trained using the training set and then evaluated using various performance metrics such as accuracy scores, R-squared scores, and test scores on the testing set. This approach allowed for the assessment of how well the machine learning models can predict the popularity of trending videos and determine which algorithm provides the best results.

5.4 Model Evaluation

The performance of the trained models was evaluated using several performance metrics. The Linear Regression model achieved a perfect score of 100% in accuracy, R-squared, and F1 score. In terms of accuracy, both the Random Forest Regressor and Decision Tree Regressor performed exceptionally well, achieving accuracy scores of 99.98%. The Support Vector Regressor was deemed unsuitable for this dataset as it generated a negative test score, which suggests that it was unable to

accurately predict the popularity of trending videos based on provided metadata.

5.5 Interpretation of the results

Based on high accuracy scores achieved by the models, we can conclude that certain factors such as the video category, likes, dislikes, and comments count have a significant impact on a video's Virality. These findings can be useful for content creators and marketers who want to optimize their content for YouTube's Trending section.

Overall, the methodology involved collecting a dataset from Kaggle, pre-processing the data, training and testing three machine learning models. The study involved evaluating the performance of different machine learning models using various metrics to gain insights into the factors that influence video Virality on YouTube. The findings of the study provide valuable insights into how metadata can be used to predict the popularity of videos on the platform. Moreover, the study underscores the importance of using machine learning algorithms to analyse large datasets and gain insights that would be difficult to obtain using traditional statistical methods. Learning techniques to analyse large datasets of metadata.

6.RESULTS

Mode	Score
Linear Regression	100%
R-squared	100%
Random Forest Regressor	99.98%
Decision Tree Regressor	99.98%

Fig.2:Results table

The results of the study show that the machine learning models achieved high

accuracy scores: a score of 100% in the Linear Regression model score and R-squared score, and a score of 99.98% in Random Forest Regressor. Additionally, the Decision Tree Regressor Test Score was 99.98%. These scores indicate that the models performed very well in predicting the Popularity of videos based on their metadata.

7.CONCLUSION

The use of machine learning techniques in analysing large datasets of metadata has shown to be a valuable tool in understanding the factors that contribute to video trendiness on YouTube. The study's findings revealed that specific factors, including the video's category, the title, the description, the view count, the comment

count, likes, dislikes, had a notable influence on a video's Virality on YouTube. Moreover, machine learning models such as the Linear Regression, Random Forest Regressor, and Decision Tree Regressor were instrumental in identifying and predicting the relationships between these factors and the popularity of trending videos were trained on the dataset, to achieving high accuracy scores. The Linear Regression model score and R-squared score both achieved a score of 100%, while the Random Forest Regressor score achieved 99.98%. Based on these results, creators and marketers can optimize their content for YouTube's Trending section by paying close attention to these factors. The findings of this study provide insights into the audience and content trends on YouTube and demonstrate the effectiveness of machine learning techniques in analysing large datasets of metadata

ORIGINALITY REPORT

8%

SIMILARITY INDEX

7%

INTERNET SOURCES

3%

PUBLICATIONS

2%

STUDENT PAPERS

PRIMARY SOURCES

1

medium.com

Internet Source

2%

2

www.frontiersin.org

Internet Source

1%

3

Zahid Halim, Sajjad Hussain, Raja Hashim Ali.
"Identifying Content Unaware Features
Influencing Popularity of Videos on YouTube:
A Study Based On Seven Regions", Expert
Systems with Applications, 2022

Publication

1%

4

www.researchgate.net

Internet Source

1%

5

dspace.daffodilvarsity.edu.bd:8080

Internet Source

1%

6

www.ijaera.org

Internet Source

1%

7

www.tandfonline.com

Internet Source

1%

8

Ekapol Wongsuparatkul, Sukree Sinthupinyo.
"View Count of Online Videos Prediction Using
Clustering View Count Patterns with
Multivariate Linear Model", Proceedings of
the 8th International Conference on
Computer and Communications
Management, 2020

Publication

<1 %

Exclude quotes On

Exclude matches Off

Exclude bibliography On