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 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 etal. (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 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.

			tike r		ORTER												
	decripte				(fishe								capital.			brasing.	
			ental)	SER			GPPRE.						U	11		Mar R prin	
ti Atjerjivov petakeran/transfeuri rjedikteti Higol			NX	16.30		Wps.) in		798	1000		Section 1					Number of Section 1877.	
escentia et rino, de tiber licosos dar urbe canadra il			1679	1639		Bjc/ju		1601	9/30		DE LINKS					application of the	
i titi i-f'olytiki i i kanji ora yake andila			FASK	FEE		Back for		1229	2900		midson					altition of the	
idika lidebularatuaraan tiidabulakute (M			F4:30	14.30		Bookley		666	10672		h5 ml					AND COURSE	
Economic del tris describé activación activativa del region de la			Hit	140	1471-	Box No.	1000	hin	56700		spofflip)					FER 9 31.4.1	
Expertises with - hint mythaglist plant in Proxitation	Wight Resile Equations		FAX	4.3	FREE	March live	3/34	56	9963	127976	lightly.	2007-010	- 2	Life	Fresh 1	COURSELL.	
lentronities by Horschild Seines will be Paide	Orbital d'Autena Senti-condition		FACE	FRE	DRE	Backley	2570	245	1580	200457	94. Sex	2007-014	. 2	Cetardes	Bylice	RESTRICTED AND	B
Period on Amplyment of Middle Colon Sealers			19.9	199	1611	Byrilly	3437	736	7000	M100	S he Desi	MITT				PROPERTY.	
interest Library Teriminat Summer baseigns made	Instructor in physics of E.L. Re-		14.3	143	193	Rendir	340	129	363	10938	linke (%)	36707		396 (m)	The Green	909×09/JAIL	8)
Sealous Hing to services, his figlicable to unchane	Personal behavioristic the		DUX	F9.50	DQ.E.	No. No.	2989	130	1381	259138	NAME .	20700	- 2	No.	Synth	Several States	8
converties to elikate ellikus unung disellaten over	Res England to blood to the experience of		NE	167	PET	Workley	117	71	600	1032	MI THE	200101	7	UH.	Destu	AND RESIDENCE.	
resoligiosidateura is tilea can emphasies ylat	this principle infrared quality		FE/S	FGS.	HEI	Box No.	2276	361	Tab.	100-10	Person	367000	- 2	491	promes.	PERSONAL PROPERTY.	e)
CONTRACTOR AND	WICHIDE WICCOMPAN		THIS	F4.80	COLD	Ballin	107.	110	197100	9750	nant nel	3007-014	1	resident.	Married	Cition of the	ĸ,
named a process contributed that are nown as drapin	Teres discharged a point point		F4:39	F9.91	F92 E	Bookie	368	117	2948	14479	ments.	30(3-1)-1	2	No. Te	State Co.	DWG STAT	E,
g bases 50 for one of the openium, i worth) to find some	Into-macrossing brows \$6 to		140	140	1473	Workline	7530	386	40	28092	though !	2007/07		No Part	1401156	Resection Act.	K.
Libetani.	for real formion, started.		FAX	FRE	DES	War his	128	52	9/25	2582	No.	367434	. 2	kauth	De Box 2	#P59/2525	
along the side of the Children to collect Constant to the	Build Enter a god accord?		FR/ST	F9.50	DRIT	Was New	1290	68	000	29509	nelicin	367-0-1	- 2	Cathol	Ownland	NO-10/15/14 T	ĸ.
r page letter/fecus primers confer in ballicular full leaves for	Christian Street age 184/5		1679	1679	1671	Sen Mar	Mi	W	100	residen	nothin.	2017/11	2	BEH	Brow Sarry	CAT Common	8
out of promit officed. The incremedy singless does in sized full Till			FA.S	FRE	res:	Sending.	200	26	1290	1000	stated.	367020		Per No.	Barrie or	458 met 2012 L	5
brancischiften Territorraperation in behand	Torrito Disetticis assertes bil		DUT	79.70	DUT	Was Not	1716	16.	MA	223670	Siller.	MEDI	- 2	to Stow	Fredret	N 000 N 15 14 11	
agreed freedom) of deadril monightest	nighters are not the		NE	NUT	HZTH	months.	2000	- 5	720	10170	Bennete	201111	. 2	h-res	RIM DE	F SPRENCE AND	
Converses a citate fractionis, the converse regions for			FEE	F9.55	FRE	Was her	268	100	205	100000	terror De	Military.	- 3	legal	Be Weit	ON THE REAL PROPERTY.	
nitral been listing Alamed II are limited to a surprise			FAX	F4.80		Balle			1170	9112	mean!	30/5/14		Sealer	Ad Serve	+Bhard H.	i
emilig, who risks this sites also his any nor habo			19.9	19/2		Water Street		SM	1967	14700		20071111				CONTRACTOR NA	
per out offer second et secondo for colorent			44	1435		Box No.		- 9	-	1200		MPOI				2062 GISCALL	
tiodrest carbod w. 2% airs to coper COX 318990				FR.ST		Barble .			00%		magent.					MACHINE HEL	
HOUSE COURSE ATTENDED TO CONTROL OF THE PARTY.	THE RESERVE OF THE PARTY OF THE	-	11	164	1104	80.704	1000	40.0				anno.			(4)	Solden	

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 Regression, Random Forest Linear 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, preprocessing 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 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	99.98%
Regressor	
Decision Tree	99.98%
Regressor	

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

ORIGINA	ALITY REPORT							
8 SIMIL	% ARITY INDEX	7 % INTERNET SOURCES	3% PUBLICATIONS	2% STUDENT PAPERS				
PRIMAF	RY SOURCES							
1	medium Internet Sour			2%				
2	2 www.frontiersin.org Internet Source							
3	"Identify Influence A Study	alim, Sajjad Hus ving Content Un ing Popularity o Based On Seven s with Applicatio	aware Feature of Videos on Yo n Regions", Ex	es ouTube:				
4	www.res	searchgate.net		1 %				
5	dspace. Internet Sour	daffodilvarsity.e	edu.bd:8080	1 %				
6	www.ija Internet Sour	_		1 %				
7	www.tai	ndfonline.com		1 %				



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

<1%

Publication

Exclude quotes

On

Exclude matches

Off

Exclude bibliography On