Dataset Nutrition Label

The Anatomy of a Trending YouTube Video

About

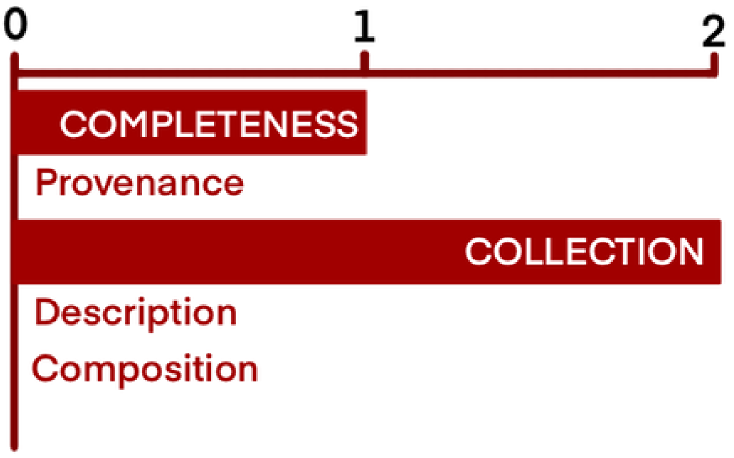
This dataset analyzes the factors that contribute to the success and popularity of YouTube videos. It examines variables such as likes, shares, saves, comments, video category, length, and posting time. It also looks at the performance of the most subscribed channels and their engagement rates across multiple platforms.

The project aims to understand the role of YouTube's algorithm in video performance, trends, and patterns in content consumption. It also explores how creators can adapt their strategies to increase engagement and remain relevant.

Data Creation Range: Aug 2020 - Dec 2021
Created By: Charity Joy Njotorahardjo, Eugene Alexander Wongso, Steven Wilbert Heng.
Content: Lists of datasets
Source: Lists of CSVs

Alert Count	* 1
Completeness	**1**
Missing	1
Provenance	**0**
Collection	**2**
Socioeconomic Bias	1
Inaccurate Prediction	1
Description	**0**
Composition	**0**

Alert Count by Category



Use Cases

1. What factors lead to video virality?
2. How does YouTube's algorithm impact video reach?
3. How can creators optimize their content for better engagement and visibility?
4. How can marketers target audiences effectively?
5. How do current events influence video popularity?
6. Can YouTube data analysis aid trend forecasting?

Badges

About Humans
Multi-source Data
Included Subpopulations
No Quality Review
No Ethical Review
Non-commercial License
Government Funded

Alert Description *

Completeness

Missing

Data is collected until December 2021. Out of 17898 rows, 96 (0.5%) of the dislikes were gone. Therefore, data removal was conducted. However, due to the large number of data, it gives almost little-to-none impact on the later analysis.

Collection

Socioeconomic Bias

Based on the data it collects to determine what is "trending," YouTube's algorithm may unintentionally favor certain creators, audiences, or types of video. Biases relating to various socioeconomic characteristics may come from this.

Inaccurate Prediction

The dataset may contain bias if the YouTube algorithm plays a substantial part in determining video popularity and trends. A prediction based on biased data may not accurately reflect the real forces behind video.

Dataset Nutrition Label

The Anatomy of a Trending YouTube Video

Dataset Informations

Information about the ongoing management of the dataset, such as
how the data will be maintained, updated, and the best contact for further inquiries.

The categories and questions that comprise this section are drawn from the terrific work of many teams, we have drawn heavily on the work of YouTube Trending Videos and 1000 Most Subscribed YouTube Channels.

Description

1. **TELL US ABOUT THIS DATASET.**
   This dataset is a set of the most subscribed YouTube Channels, the engagement they get, and trending videos classified by date and category.

2. **IS THERE AN INTENDED PURPOSE FOR THE DATASET? WHAT DOMAIN WAS IT DESIGNED FOR?**
   The purpose of this dataset is to assess the engagement of trending YouTube videos and the most subscribed channels on YouTube.

3. **WHAT IS THE LICENSE UNDER WHICH THE DATASET IS MADE AVAILABLE?**
   Public Domain

4. **ARE THERE TASKS FOR WHICH THE DATASET WOULD BE CAUTIONED AGAINST BEING USED? IF SO, PLEASE PROVIDE A DESCRIPTION.**

No

5. **HAS THE DATASET BEEN USED FOR ANY TASKS ALREADY? IF SO, PLEASE PROVIDE A DESCRIPTION AND LINKS TO PAPERS OR SYSTEMS USING THE DATASET**

No

Composition

1. **DOES THE DATASET HAVE A METADATA REPOSITORY OR DATA DICTIONARY? IF YES, PLEASE PROVIDE THE LINK AND IF NOT, PLEASE EXPLAIN WHAT EACH FIELD MEANS.**

Yes,
https://github.com/INFO-201-Fall-2023/final-project-repositories-eugenewongso/settings

2. **WHAT DATA DOES EACH INSTANCE CONSIST OF? FOR EXAMPLE, DOES IT CONSIST OF RAW DATA (E.G., UNPROCESSED TEXT OR IMAGES) OR FEATURES? IN EITHER CASE, PLEASE PROVIDE A DESCRIPTION.**

No images or unprocessed text in this dataset.

3. **IS YOUR DATASET A SAMPLE? IF SO, WHAT WAS THE SAMPLING STRATEGY USED (E.G. DETERMINISTIC, PROBABILISTIC WITH SPECIFIC SAMPLING PROBABILITIES), AND DOES IT ACCURATELY REPRESENTS THE INTENDED OUTPUT?**

Dataset is not a sample.

4. **WAS ANY PREPROCESSING/CLEANING/LABELING OF THE DATA DONE? IF SO, PLEASE PROVIDE A DESCRIPTION.**

   Yes, data cleaning and augmentation is done through RStudio.

5. **ARE THERE ANY ERRORS, SOURCES OF NOISE, OR REDUNDANCIES IN THE DATASET? IF SO, PLEASE PROVIDE A DESCRIPTION**

   In the process of collecting the data, a very few (0.5%) of them had no data regarding the amount of likes and dislikes. Therefore, data removal was conducted.

6. **IS THERE CONFIDENTIAL DATA INCLUDED IN THIS DATASET?**

   No.

7. **DOES THE DATASET IDENTIFY ANY SUBPOPULATIONS? IF YES, PLEASE INDICATE THEIR DISTRIBUTIONS WITHIN THE DATASET WHERE POSSIBLE / AVAILABLE.**

   Yes, dataset consists of video categories which is considered as subpopulation in the context of this study –– analyzing video performance and trends.

Provenance

1. **WHO CREATED THE DATASET AND ON BEHALF OF WHICH ENTITY?**

   Rishav Sharma, an Individual researcher.

2. **HOW IS / WAS THE COLLECTION AND MANAGEMENT OF THIS DATASET FUNDED?**

   N/A, higher chance of it being personally funded.

Collection

1. **OVER WHAT TIMEFRAME WAS THE DATA COLLECTED? IF THE DATA CONTINUES TO BE UPDATED, PLEASE INDICATE.**
   The dataset is updated daily, but the data used in this study is as of December 2021 (from August 2020).

2. **WHAT MECHANISMS OR PROCEDURES WERE USED TO COLLECT THE DATA?**
   This data was collected using the YouTube API.

3. **IF INDIVIDUALS' DATA IS INCLUDED IN THIS DATASET, DID THOSE INDIVIDUALS CONSENT TO THE COLLECTION AND USE OF THEIR DATA? IF SO, PLEASE DESCRIBE CONSENT PROCEDURE.**
   N/A.

4. **IF INDIVIDUALS' DATA IS INCLUDED IN THIS DATASET, WAS THIS DATA ALTERED TO ENSURE HIGHER LEVELS OF PRIVACY? IF YES, PLEASE DESCRIBE ANY PRIVACY PROCEDURES FOLLOWED WITH REGARDS TO THIS DATASET (ANONYMIZATION EFFORTS, PRIVACY PROTOCOLS, SUPRESSION TECHNIQUES, ETC).**
   N/A.

5. **HAS THE DATA BEEN REVIEWED FOR QUALITY?**

No.

6. **WERE ANY ETHICAL REVIEW PROCESSES CONDUCTED (E.G., BY AN INSTITUTIONAL REVIEW BOARD)? IF SO, PLEASE PROVIDE A DESCRIPTION OF THESE REVIEW PROCESSES, INCLUDING THE OUTCOMES, AS WELL AS A LINK OR OTHER ACCESS POINT TO ANY SUPPORTING DOCUMENTATION.**

No.

Management

1. **HOW IS THE DATA BEING MANAGED AT REST AND IN TRANSIT?**

Internally.

2. **HOW CAN THE DATA BE ACCESSED? IF THIS IS GOING TO CHANGE OVER TIME, PLEASE INDICATE THIS.**

It is accessible through Kaggle. A subsidiary of Google LLC, which is also an online community of data scientists and machine learning practitioners. No information whether this is going to change overtime or not.

3. **WILL THE DATASET BE UPDATED (E.G., TO CORRECT LABELING ERRORS, ADD NEW INSTANCES, DELETE INSTANCES)? IF SO, PLEASE DESCRIBE HOW OFTEN, BY WHOM, AND HOW UPDATES WILL BE COMMUNICATED TO USERS (E.G., MAILING LIST, GITHUB)?**

Dataset is updated daily, but for this study, it is only going to cover data from August 2020 up until December 2021.