DATA_Sheet

Motivation

- **Purpose**: The dataset was created to analyze, categorize, and gain insights into YouTube channels for enhancing strategic decision-making in marketing, content strategy, and partnerships.
- **Creators and Funders**: The dataset, sourced from Kaggle, was originally created by Nidula Elgiriyewithana. It has since been augmented with channel descriptions and primary languages using large language models (ChatGPT-4, Google Gemini, META LLMA 2).

Composition

- **Data Representation**: The dataset comprises information about YouTube channels, including metrics like video views, subscriber counts, and other channel-specific information.
- **Instance Count**: Approximately 1000.
- **Missing Data**: Depending on the type of information, up to 73 lines are missing.
- **Confidentiality**: The data is public.

Collection Process

- **Data Acquisition**: Data was acquired from public sources and aggregated into the Kaggle dataset.
- **Sampling Strategy**: The dataset represents the top approximately 1000 YouTube channels by subscribers.
- **Collection Timeframe**: Throughout 2023.

Preprocessing/Cleaning/Labeling

- **Preprocessing Details**: The dataset includes preprocessed elements such as channel descriptions enhanced with language models.
- **Raw Data Preservation**: Data is preserved in a separate file.

Uses

- **Potential Uses**: Beyond analyzing YouTube channel performance, the dataset could be used for trend analysis, market research, or training other machine learning models for content prediction.
- **Composition and Collection Considerations**: The dataset's utility might be influenced by its focus on top channels, which may not be representative of all types of content on YouTube. Future users should consider this when generalizing findings or applying insights to broader contexts.
- **Inappropriate Uses**: The dataset should not be used for decisions affecting individual privacy or for constructing personalized recommendations without consent.

Distribution

- **Current Distribution**: Distributed via Kaggle.
- **Copyright/IP**: The dataset is subject to Kaggle’s standard terms of use.

Maintenance

- **Frequency**: Updated annually.
- **Maintenance Responsibility**: Maintained by the dataset owner.