Project Proposal

Yujie Gong, Mo Wang, Chengxi Zhou

Project Title and Team Members

Project Title: YouTube Healthcare Dashboard: Engagement and Content Analysis Across Health Education Topics

Team Members:
Yujie Gong, Mo Wang, Chengxi Zhou

Research Question

How do viewers' attitudes and engagement differ across different health education topics on YouTube, and what keywords or content characteristics make healthcare videos more appealing to the public?

YouTube has become one of the most widely used platforms for learning about health, wellness, and science-related topics. Its accessibility, visual format, and recommendation algorithms make it a convenient source of information for diverse audiences.

However, public reliance on YouTube and the ways health information is created and consumed, vary widely across topics. Some areas, such as mental health or fitness, attract high engagement from influencers and non-professional creators, while more specialized medical topics are less represented or primarily produced by professional organizations.

These differences in content quality, creator type, and audience interest lead to uneven patterns in how people use YouTube for health education, highlighting the need to examine which topics gain attention and how viewers interact with them.

Existing Work

YouTube has become one of the most influential platforms for health communication, allowing the public to access medical information in a highly accessible and engaging format. Many individuals turn to YouTube to learn about diseases, treatments, and preventive practices, which has made it a crucial medium for health education. However, the credibility of YouTube health content remains inconsistent, as studies have documented both accurate educational materials and misleading or anecdotal information (Madathil et al. 2015). This dual nature highlights both the educational potential and the public health risks of YouTube as a source of health information.

A substantial body of topic-specific research has examined how health information is presented and received across different medical areas. For example, studies on chronic diseases such as diabetes and hypertension have found that physician-created videos generally provide more reliable information, whereas patient-generated videos tend to attract higher engagement and audience interaction (Sood et al. 2011). Research on cancer-related topics, including breast cancer, prostate cancer, and oral cancer, shows that many videos contain partially accurate or incomplete information, often prioritizing personal narratives or emotional storytelling over factual precision (Gulve et al. 2022). Similar patterns have been observed during infectious disease outbreaks, such as COVID-19, where YouTube became a major communication channel but also a source of widespread misinformation (Li et al., 2020). In the mental-health domain, studies of depression-related videos highlight that non-professional creators often shape illness narratives and stigma perceptions within the platform (Devendorf, Bender, and Rottenberg 2020). Collectively, these findings demonstrate YouTube's significant role in shaping health understanding, yet most prior work remains focused on single disease areas rather than comparing engagement across broader health-education themes.

Beyond topic-specific findings, prior research has also demonstrated practical approaches for collecting and analyzing YouTube data. Carvache-Franco et al. used the YouTube Data API to retrieve large-scale comment datasets and applied word-association methods to identify major discussion themes(Carvache-Franco et al. 2023). Their study provides a concrete example of how API-based workflows can be used to examine public attitudes at scale. Complementing this methodological perspective, Devendorf et al. (Devendorf, Bender, and Rottenberg 2020)conducted a detailed content analysis of depression-related videos, illustrating how content characteristics—such as framing of causes, treatments, and illness course—shape public perceptions and patterns of engagement. Together, these studies offer both a technical foundation for API-driven data collection and a thematic understanding of how health information is represented on YouTube. These insights directly inform our project, which extends prior work by comparing viewer engagement across different categories of health-education content rather than focusing on a single medical topic.

Outline

Our project aims to analyze how viewers' attitudes and engagement differ across various health education topics on YouTube and to identify which keywords or content characteristics make healthcare videos more appealing to the public. To achieve this, we will follow the steps below.

1. Data Collection

We will use the YouTube Data API to extract metadata and comments from videos under different health-related topics (e.g., diabetes, mental health, vaccination, nutrition). The data collected includes video titles, descriptions, publication dates, view counts, like counts, comment counts, and a subset of user comments for sentiment analysis.

2. Data Cleaning and Preprocessing

We will perform data cleaning using tidyverse tools to remove duplicates, handle missing values, and standardize textual data (e.g., lowercasing, removing stopwords, and stemming). We will also categorize videos by topic using both keyword-based filtering and, if needed, simple text classification.

3. Sentiment and Text Analysis

We will use natural language processing (NLP) techniques to analyze the tone of viewer comments and extract common keywords and phrases. Word frequency analysis, sentiment scoring, and topic modeling will be employed to understand viewers' attitudes toward different health topics.

4. Engagement Analysis

We will investigate relationships between views, likes, comments and content characteristics such as topic, title keywords, video length, upload date.

5. Visualization and Reporting

We will visualize key findings through interactive plots and summary tables, highlighting differences across topics. The final data analytic product will be a dash board summarizing our analysis, including data visualizations and discussion of insights. # Packages and Software

Our programming environment will primarily be R using Quarto for documentation and report generation. The main packages and tools include: Data Cleaning and Manipulation: tidyverse (including dplyr, stringr, and tidyr) for data wrangling. Text and Sentiment Analysis: tidytext and textdata for tokenization, stopword removal, and sentiment scoring. Visualization and Reporting: ggplot2 for data visualization. And wordcloud for displaying frequent keywords.

Accessing the Data

We will collect data through the publicly available YouTube Data API v3, which provides open access to structured metadata about videos, channels, and user engagement on YouTube. This API allows researchers to retrieve information such as video titles, publication dates, channel names, view counts, likes, comment counts, and topic tags for any publicly accessible video. Because the API is maintained by Google and does not require access to private user data, it is a reliable and ethical source for studying online health communication patterns. The dataset we aim to build will include videos across a range of health education topics, for example, mental health awareness, nutrition, disease prevention, and exercise for wellbeing. For each topic, we plan to retrieve a large sample of videos that meet the search criteria and compile their descriptive and engagement statistics. This will create a structured dataset that reflects how audiences interact with health-related content online.

To collect the data, we will query the YouTube API by submitting topic keywords and retrieving results in multiple batches to capture a sufficient number of videos (up to 1,000 per topic). Each query returns a JSON-formatted dataset containing both video metadata and engagement measures, which we will process and combine into a single dataset for analysis.

Dataset Description

Variable	Description
video_id	Unique identifier for each YouTube video
title	Title of the video as displayed on YouTube
channel_name	Name of the channel that published the video
published_date	Date the video was uploaded
views	Total number of times the video has been viewed
likes	Number of likes received
comments	Number of comments posted by viewers

Variable	Description
tags	Keywords or tags assigned by the video creator
duration	Length of the video (in seconds or minutes)
topic_category	Health topic associated with the video (e.g., mental health, fitness)

Example of Data Access

To demonstrate our ability to access and collect data through the YouTube Data API, we created a small test repository on GitHub that includes an example of extracting mental health–related videos from YouTube and converting them into a structured dataset.

Api-testing.qmd — an example showing how to query the YouTube Data API and convert the output into a structured dataset. **youtube_mental_health_1000.csv** — a sample dataset containing 1,000 videos retrieved using the keyword *"mental health awareness."*

Repository links:
- GitHub Repository
- youtube_mental_health_1000.csv

Programming Paradigms

Our project will combine the functional programming and machine learning paradigms to collect, process, and analyze YouTube data related to health education topics. The functional programming paradigm will be used for data collection and cleaning, taking advantage of R packages such as dplyr, purrr, and stringr. This approach allows us to transform raw JSON outputs from the YouTube Data API, filtering, mapping, and summarizing information such as video titles, views, likes, and tags. Functional programming ensures code simplicity, readability, and reproducibility, which are essential for working with large and continuously updated datasets.

The machine learning paradigm will then be applied to identify patterns and predict engagement levels across different health education topics. Using algorithms such as multiple regression, decision trees, or clustering, we will model how factors like video duration, topic category, title length, and keyword frequency relate to engagement metrics such as views, likes, and comments. This paradigm allows the project to move beyond descriptive statistics toward predictive insights, helping us understand what makes certain healthcare videos more appealing to the public.

We also will possibly use the object-oriented programming paradigm in our project. The object-oriented programming paradigm will be used to structure our workflow into reusable ways for interacting with the YouTube Data API. We plan to develop an API handler object that includes key elements such as the API credentials, search parameters, and methods for retrieving, processing, and saving video data. This approach allows us to easily adapt the same framework to different health-related topics, such as nutrition, mental health, or exercise, without duplicating code.

Planned Data Analytic Product

We plan to develop an interactive dashboard to summarize and visualize the insights derived from our YouTube healthcare analysis. This dashboard will serve as the central product of our project, enabling users to explore and understand the differences in viewer engagement and content characteristics across various health education topics.

Built using the R Shiny framework, the dashboard will feature multiple linked visualizations. Key components will include:

1. Topic Comparison Panel: This section will allow users to select and compare different health topics. It will display metrics such as average views, likes, and comment counts, providing a high-level overview of engagement patterns.
2. Content Characteristics Explorer: Here, we will visualize the keywords and tags that are most frequently associated with highly-engaged videos within each topic. This could be presented through interactive word clouds or bar charts, helping to answer what makes certain health videos more appealing.
3. Sentiment & Engagement Correlation View: If comment sentiment analysis is performed, this visualization will illustrate the relationship between the sentiment of viewer comments and the video's engagement metrics, revealing how audience attitudes correlate with popularity.

The dashboard will transform our structured dataset into a dynamic and accessible tool. It will allow interactive exploration and make it easier to understand how audience attitudes and engagement differ across health-education topics on YouTube.

Timeline

Week 1:

- Test access to YouTube Data API and write initial API query script.
- Retrieve a small sample dataset (eg., 50–100 videos per topic) to confirm workflow.

Week 2 :

- Collect full datasets for each selected health topic.
- Begin cleaning and organizing the dataset.
- Conduct preliminary descriptive summaries and check for data quality issues.

Week 3 :

- Perform main analyses: topic comparisons, keyword/tag exploration, and basic sentiment analysis.
- Start building visualizations to be used in the dashboard.

Week 4 :

- Build the Shiny dashboard and integrate visualizations.
- Finalize the project report, polish code, and check reproducibility.
- Prepare final submission.

Team Responsibilities

To ensure an efficient workflow, our project tasks will be divided among the three team members as follows:

Yujie

- Lead the data extraction process using the YouTube Data API.

- Develop and test API query scripts.

- Oversee data cleaning and preprocessing workflows.

- Contribute to dashboard development and final report writing.

Mo

- Lead the text and sentiment analysis using tidytext and related NLP tools.
- Conduct keyword, tag, and topic exploration across health topics.
- Assist with engagement analysis and data visualization development.
- Contribute to dashboard integration.

Chengxi

- Lead statistical and machine learning modeling (regression, clustering, predictive analyses).
- Analyze relationships between engagement metrics and video characteristics.
- Build interactive visualizations for the Shiny dashboard.
- Support reproducibility checks and preparation of the final submission.

All team members will collaboratively participate in: - Designing the dashboard layout - Writing the final report and presentation - Reviewing code, ensuring reproducibility, and preparing the final presentation

Carvache-Franco, O., M. Carvache-Franco, W. Carvache-Franco, and O. Martin-Moreno. 2023. "Topics and Destinations in Comments on YouTube Tourism Videos During the Covid-19 Pandemic." *PLOS ONE* 18 (3): e0281100. https://doi.org/10.1371/journal.pone.0281100.

Devendorf, A., A. Bender, and J. Rottenberg. 2020. "Depression Presentations, Stigma, and Mental Health Literacy: A Critical Review and YouTube Content Analysis." *Clinical Psychology Review* 78: 101843. https://doi.org/10.1016/j.cpr.2020.101843.

Gulve, N. D., P. R. Tripathi, S. D. Dahivelkar, M. N. Gulve, R. N. Gulve, and S. J. Kolhe. 2022. "Evaluation of YouTube Videos as a Source of Information about Oral Self-Examination to Detect Oral Cancer and Precancerous Lesions." *Journal of International Society of Preventive & Community Dentistry* 12 (2): 226–34. https://doi.org/10.4103/jispcd.JISPCD_277_21.

Madathil, K. C., A. J. Rivera-Rodriguez, J. S. Greenstein, and A. K. Gramopadhye. 2015. "Healthcare Information on YouTube: A Systematic Review." *Health Informatics Journal* 21 (3): 173–94. https://doi.org/10.1177/1460458213512220.

Sood, A., S. Sarangi, A. Pandey, and K. Murugiah. 2011. "YouTube as a Source of Information on Kidney Stone Disease." *Urology* 77 (3): 558–62. https://doi.org/10.1016/j.urology.2010.07.536.