YouTube Healthcare Dashboard: Engagement, Sentiment, and Popularity in Online Health Education

Team Members: Yujie Gong, Mo Wang, Chengxi Zhou
Final Product: Interactive R Shiny Dashboard

1. Overview of the Project

YouTube has become one of the most widely used platforms for learning about health, wellness, and medical science. Its accessibility, visual format, and recommendation algorithms make it an appealing source of information for diverse audiences. However, the quality, credibility, and engagement patterns of health-related content vary substantially across topics and creator types. While some videos are produced by professional organizations or medical experts, others are created by influencers or non-professional creators who may prioritize storytelling or emotional appeal over accuracy.

This project investigates how viewers' attitudes and engagement differ across health education topics on YouTube and identifies which content characteristics are associated with higher public engagement. Building on our original proposal, we developed an interactive Shiny dashboard that integrates large-scale YouTube data, sentiment analysis, and statistical modeling to provide an accessible and reproducible analytic tool.

2. Research Question

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

Specifically, we examine differences between:

- **Professional medical topics:** cardiovascular diseases, diabetes, and antidepressants
- **Influencer / wellness topics:** probiotics, hormone balance, and anti-aging

3. Existing Work

Prior research has established YouTube as a major platform for health communication, but with inconsistent credibility. Studies have documented that physician-created videos often provide higher-quality information, while patient- or influencer-generated videos tend to receive higher engagement (Madathil et al., 2015; Sood et al., 2011). Topic-specific research on cancer, infectious diseases, and mental health further shows that emotional framing and personal narratives frequently drive attention, sometimes at the expense of accuracy (Gulve et al., 2022; Devendorf et al., 2020).

Methodologically, previous work has demonstrated the feasibility of using the YouTube Data API for large-scale analysis of engagement and comments, including keyword and sentiment-based approaches (Carvache-Franco et al., 2023). However, most existing studies focus on a single disease area and rely on static analyses.

Our project extends this literature by:

1. Comparing engagement patterns across multiple health education topics.
2. Integrating sentiment analysis of titles and descriptions.
3. Providing an interactive dashboard that allows users to explore results dynamically rather than through static summaries.

4. Challenges & Accomplishments

During this project, we encountered several practical challenges and achieved the following accomplishments:

1. **Working with Large-Scale API Data**
   One major challenge was handling data collected from the YouTube API. The data came with heterogeneous formats, missing values, and varying levels of completeness across videos and topics. Additionally, managing large volumes of video-level data required careful preprocessing to ensure the dashboard remained responsive and efficient.
2. **Designing an Interactive and Scalable Dashboard**
   Building a multi-tab interactive dashboard posed challenges in balancing functionality and usability. We needed to design filters (topics, creator type, minimum views) that allowed meaningful exploration while avoiding overly complex or confusing user interactions. Ensuring that plots updated smoothly as filters changed was a key technical and design challenge.
3. **Integrating Text-Based Sentiment Analysis**
   Applying sentiment analysis to video titles and descriptions introduced additional complexity, as sentiment scores can be noisy and vary widely across topics. Interpreting these scores in a meaningful way—especially when relating sentiment to video popularity—required careful visualization and cautious interpretation.
4. **Incorporating Model-Based and Machine Learning Components**
   Integrating statistical modeling and machine learning methods, such as regression analysis and clustering-based insights, was another challenge. We needed to ensure that these methods complemented the exploratory dashboard rather than overwhelming users, while still providing interpretable results about what factors drive video popularity.

5. Shiny Dashboard Overview

The final analytic product is an interactive Shiny dashboard with the following components:

1. **Home:** Project description and data overview.
2. **Overview:** Summary tables and distributions of views by topic and creator type.
3. **Engagement:** Like rate vs. comment rate scatterplots.
4. **Sentiment:** Sentiment summaries and sentiment–views relationships.
5. **Topic Comparison:** Average views across topics.
6. **Video Table:** Ranked table of top-performing videos.
7. **Drivers of Views:** Regression results with effect sizes and confidence intervals.

All views are dynamically linked to user-selected filters.

6. Usability and Documentation

The dashboard prioritizes accessibility and interpretability. Statistical results are presented through intuitive visualizations, and all code is thoroughly documented to ensure reproducibility. The application is suitable for both exploratory analysis and instructional use, and it is designed to be easily navigable for users with limited statistical background.

7. Originality and complexity

From a programming perspective, the project involves complex data processing pipelines, including handling high-volume, highly skewed engagement data, integrating multiple analytical components, and dynamically updating outputs through a Shiny interface. The dashboard design required careful coordination between backend data transformations and frontend interactivity, as well as thoughtful consideration of usability for non-technical users.

The project's creativity aspect is our idea of building a comparative, cross-topic focus dashboard. By transforming raw YouTube data into an interactive, user-centered analytic tool, the project moves beyond standard descriptive analysis and demonstrates a creative application of statistical thinking and software design principles.

8. Key Findings

- **Likes are the strongest predictor of view count**, while topic category and creator type contribute comparatively little once engagement is accounted for.

- **Engagement efficiency varies substantially by topic.**
  Probiotics and cardiovascular-related videos demonstrate higher engagement relative to their view counts, whereas anti-aging videos tend to attract high viewership but only moderate engagement.

- **Extremely high-view videos are rare**, and the resulting right-skewed distribution limits how much filtering can be applied before the statistical model suffers from insufficient usable data.

9. Lessons Learned

Through this project, we learned the importance of adopting a user-centered mindset when designing data products. Rather than focusing solely on our own analytic interests, we considered what potential users might find intuitive, useful, and engaging, which significantly influenced the structure and functionality of the dashboard. We also gained experience in integrating data processing with visualization, learning how backend analytical decisions directly shape frontend communication. This project strengthened our understanding of full-stack product design, including the interaction between user interface choices and data workflows.

We also developed practical experience with collaborative software development using GitHub. By regularly pushing and pulling code, managing branches, and resolving merge conflicts, we learned how to coordinate contributions efficiently within a shared repository. This workflow was essential for maintaining reproducibility, avoiding code conflicts, and enabling smooth collaboration across team members.

Finally, we deepened our familiarity with the YouTube Data API and became aware of the wide range of open APIs provided by Google, providing us with more opportunities for future exploration and extension of data-driven applications.

10. Limitations

Despite providing useful insights into engagement patterns and sentiment trends in health-related YouTube content, this project has several limitations that should be acknowledged.

First, data availability is constrained by YouTube API restrictions, which limit the number of retrievable videos and may prevent the dataset from fully capturing the breadth of healthcare-related content on the platform. Additionally, some health topics are represented by substantially fewer videos than others, which may introduce imbalance and bias into cross-topic comparisons. As a result, observed differences across topics should be interpreted with caution.

Second, commonly used engagement metrics such as views and likes primarily reflect audience interest and visibility rather than content quality or medical accuracy. High engagement does not necessarily indicate reliable or evidence-based health information, which limits the extent to which engagement can be used as a proxy for public health impact. Third, sentiment analysis in this project is based on YouTube video titles and descriptions, which often contain sarcasm, promotional language, or noise. These linguistic characteristics can reduce the accuracy of sentiment scores and may not fully reflect viewers' true attitudes or emotional responses to the content. Consequently, sentiment measures should be interpreted as approximate indicators rather than precise representations of audience sentiment.

Finally, the dashboard and statistical analyses identify associations between content features and engagement outcomes but do not support causal inference. Relationships observed in the analysis may be influenced by unmeasured confounding factors, and causal conclusions about the effects of sentiment or content characteristics on engagement cannot be drawn.

11. Future Directions

This project establishes a foundation for understanding engagement and sentiment patterns in online health-related video content, but several extensions could strengthen its analytical depth and practical relevance.

Future work could incorporate more advanced natural language processing models, such as BERT-based or transformer-based sentiment classifiers, to better capture nuanced emotional tones, sarcasm, and contextual meaning in video metadata. Expanding textual analysis to include video transcripts or user comments could also provide a more comprehensive view of audience perception and interaction.

Future projects could also use data sources beyond YouTube, such as TikTok, Reddit, or health-focused forums, which would allow for cross-platform comparisons and improve the generalizability of findings. Including external indicators of content credibility, such as expert review scores or references to authoritative health sources, could further distinguish popularity from informational quality.

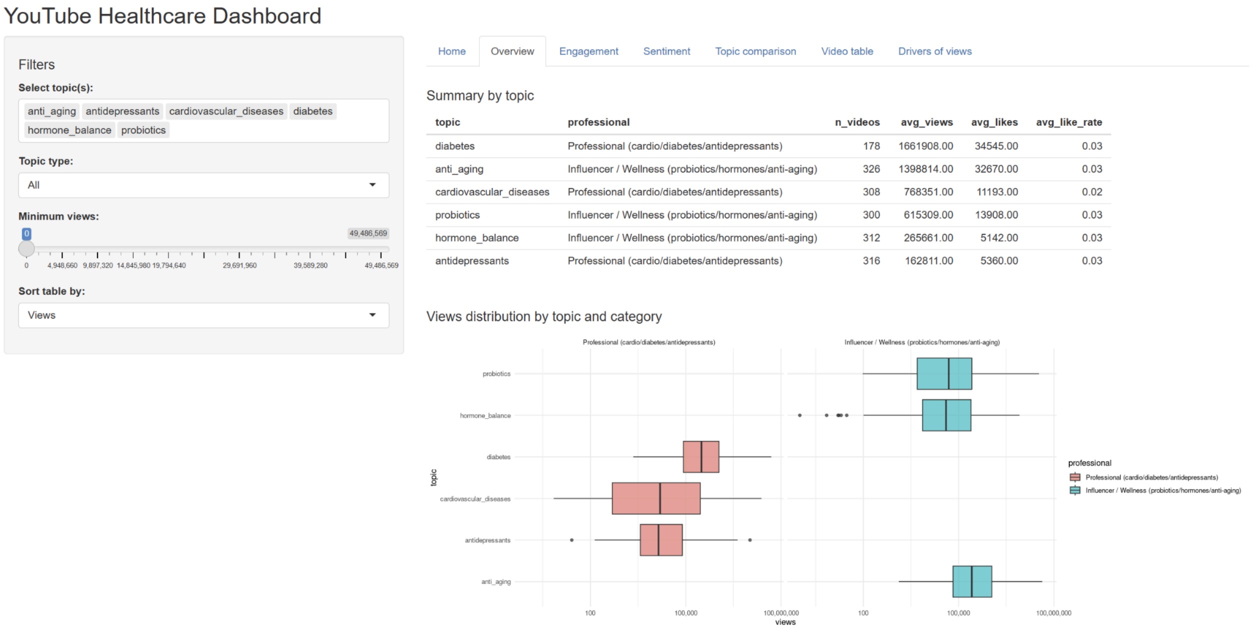
From an analytical perspective, future studies could explore causal or quasi-experimental methods, such as matched comparisons or longitudinal designs, to better assess how content features influence engagement over time. Enhancing dashboard functionality with richer visualizations or real-time updates may also improve its usefulness for researchers, content creators, and public health practitioners seeking to monitor and respond to emerging health communication trends.

12. Conclusion

This project demonstrates how large-scale YouTube data, sentiment analysis, and statistical modeling can be combined into an interactive and accessible analytic tool. By comparing engagement across professional and influencer-driven health content, the YouTube Healthcare Dashboard provides meaningful insights into how online health information is consumed and what characteristics drive public attention.

For more demonstrations of our product:
Link for dashboard: https://flusightcategoricalprediction.shinyapps.io/healthcare_dashboard/
Link for our codes: https://github.com/jhu-statprogramming-fall-2025/project04-vivo-50

Below we also attach some plots from the demo to show its functionalities:



YouTube Healthcare Dashboard

Filters

Select topic(s): anti_aging antidepressants cardiovascular_diseases diabetes hormone_balance probiotics

Topic type: All

Minimum views: 0 – 49,486,569

Sort table by: Views

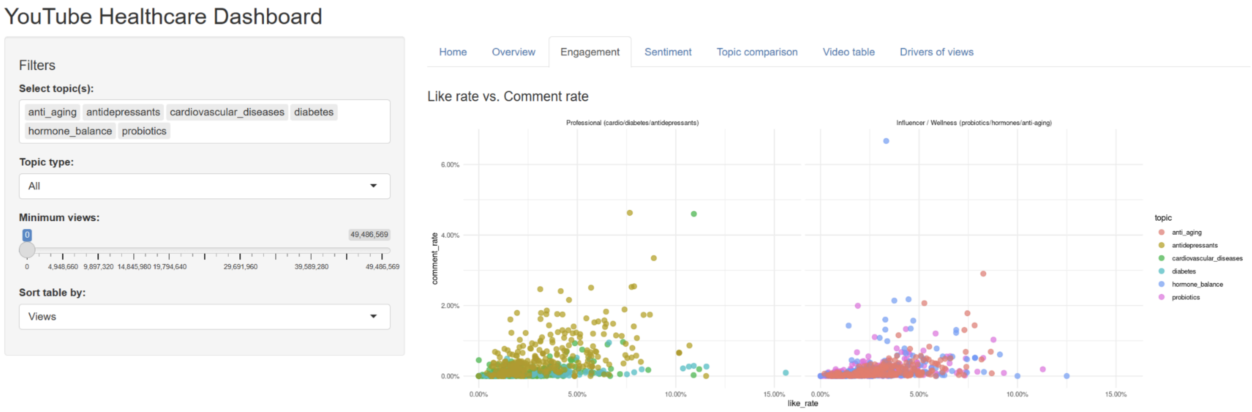
Summary by topic

topic	professional	n_videos	avg_views	avg_likes	avg_like_rate
diabetes	Professional (cardio/diabetes/antidepressants)	178	1661908.00	34545.00	0.03
anti_aging	Influencer / Wellness (probiotics/hormones/anti-aging)	326	1398814.00	32670.00	0.03
cardiovascular_diseases	Professional (cardio/diabetes/antidepressants)	308	768351.00	11193.00	0.02
probiotics	Influencer / Wellness (probiotics/hormones/anti-aging)	300	615309.00	13908.00	0.03
hormone_balance	Influencer / Wellness (probiotics/hormones/anti-aging)	312	265661.00	5142.00	0.03
antidepressants	Professional (cardio/diabetes/antidepressants)	316	162811.00	5360.00	0.03

Views distribution by topic and category

Professional (cardio/diabetes/antidepressants)
Influencer / Wellness (probiotics/hormones/anti-aging)
topic
probiotics
hormone_balance
diabetes
cardiovascular_diseases
antidepressants
anti_aging
100
100,000
100,000,000
100
100,000
100,000,000
views
professional
Professional (cardio/diabetes/antidepressants)
Influencer / Wellness (probiotics/hormones/anti-aging)

Figure 1. Overview summary table and view distribution by topic and creator type.
This figure summarizes video counts, average views, and likes across topics. Boxplots show the distribution of view counts by topic, highlighting differences between Professional and Influencer/Wellness creators.



YouTube Healthcare Dashboard

Filters

Select topic(s): anti_aging antidepressants cardiovascular_diseases diabetes hormone_balance probiotics

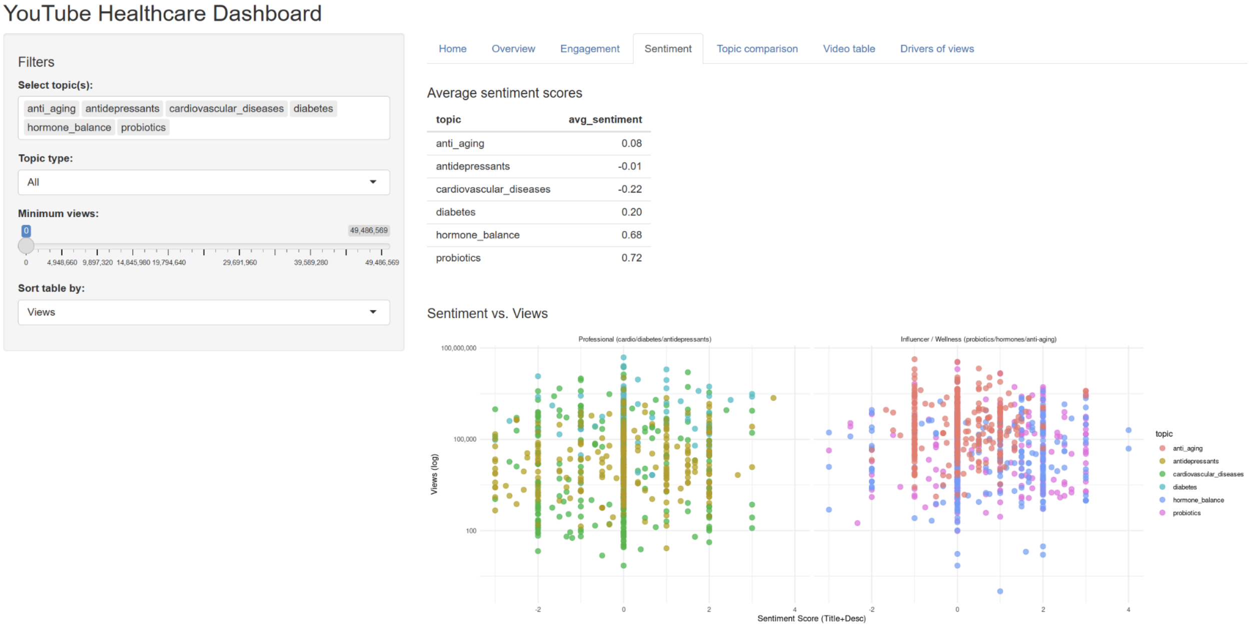
Topic type: All

Minimum views: 0 – 49,486,569

Sort table by: Views

Like rate vs. Comment rate

Figure 2. Like rate versus comment rate by topic and creator type.
This scatter plot compares engagement patterns across Professional and Influencer/Wellness creators. While professional videos tend to receive relatively higher comment rates at lower view thresholds, influencer videos generally show higher like rates and comment rates as the minimum view threshold increases, indicating differences in engagement efficiency across creator types.



YouTube Healthcare Dashboard

Filters

Select topic(s): anti_aging antidepressants cardiovascular_diseases diabetes hormone_balance probiotics

Topic type: All

Minimum views: 0 49,486,569

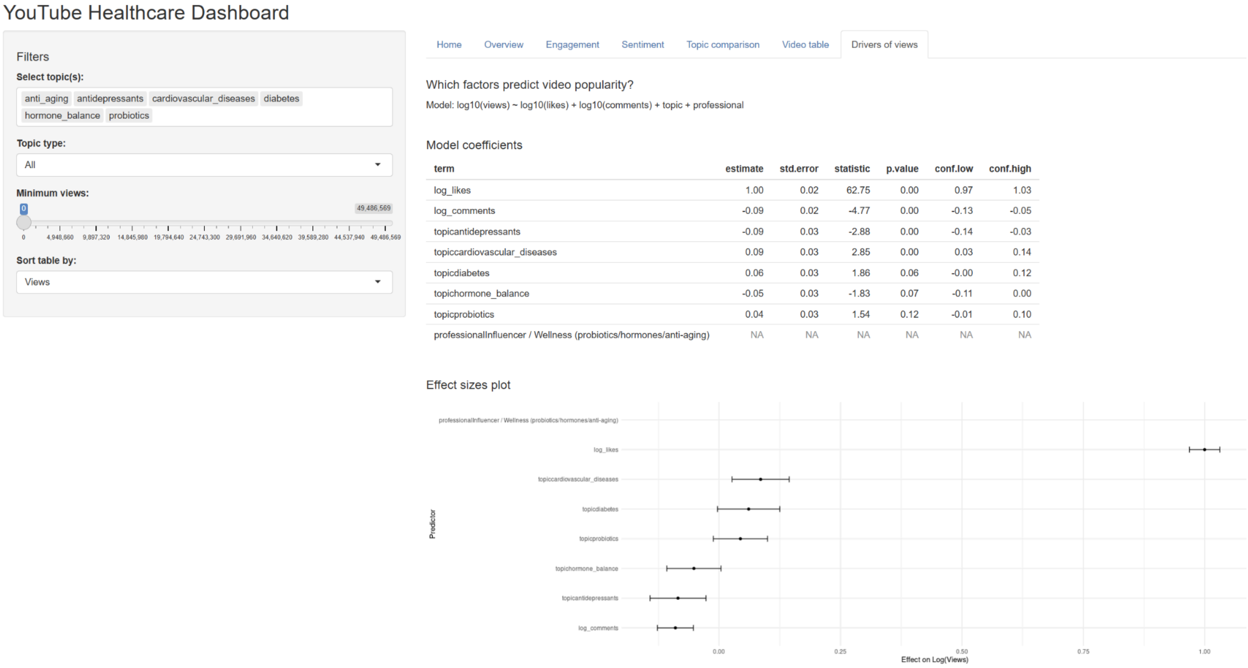
Sort table by: Views

Average sentiment scores

topic	avg_sentiment
anti_aging	0.08
antidepressants	-0.01
cardiovascular_diseases	-0.22
diabetes	0.20
hormone_balance	0.68
probiotics	0.72

Sentiment vs. Views

Figure 3. Sentiment patterns across topics and their relationship with views.
The table summarizes average sentiment scores by topic, while the scatter plots illustrate the relationship between sentiment (derived from video titles and descriptions) and view counts for Professional and Influencer/Wellness creators. Although sentiment levels vary substantially across topics, the visualizations suggest that sentiment is only weakly related to video popularity, with no clear trend indicating that more positive sentiment consistently leads to higher view counts.



YouTube Healthcare Dashboard

Filters

Select topic(s): anti_aging antidepressants cardiovascular_diseases diabetes hormone_balance probiotics

Topic type: All

Minimum views: 0 49,486,569

Sort table by: Views

Which factors predict video popularity?

Model: log10(views) ~ log10(likes) + log10(comments) + topic + professional

Model coefficients

term	estimate	std.error	statistic	p.value	conf.low	conf.high
log_likes	1.00	0.02	62.75	0.00	0.97	1.03
log_comments	-0.09	0.02	-4.77	0.00	-0.13	-0.05
topicantidepressants	-0.09	0.03	-2.88	0.00	-0.14	-0.03
topiccardiovascular_diseases	0.09	0.03	2.85	0.00	0.03	0.14
topicdiabetes	0.06	0.03	1.86	0.06	-0.00	0.12
topichormone_balance	-0.05	0.03	-1.83	0.07	-0.11	0.00
topicprobiotics	0.04	0.03	1.54	0.12	-0.01	0.10
professionalInfluencer / Wellness (probiotics/hormones/anti-aging)	NA	NA	NA	NA	NA	NA

Effect sizes plot

Figure 4. Drivers of video popularity based on a linear regression model.
This figure presents the estimated coefficients and effect sizes from a linear regression model predicting log-transformed view counts using engagement metrics (likes and comments), topic

indicators, and creator type. The effect size plot shows that likes are the strongest predictor of view count, with substantially larger and more stable effects than topic or creator type, which contribute comparatively less to explaining video popularity.