MEDINFO 23
8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

YouTube Videos for Public Health Literacy?

A Machine Learning Pipeline to Curate Covid-19 Videos

Yawen Guo, MS [1], Xiao Liu, PhD [2], Anjana Susarla, PhD [3], Rema Padman , PhD[4]

[1] *University of California, Irvine*
[2] *Arizona State University*
[3] *Michigan State University*
[4] *Carnegie Mellon University*

Funding: NIH/National Library of Medicine #R01LM013443

Health Literacy and Covid-19 Self-care

- Global burden of disease – “perfect storm of rising chronic diseases and public health failures fueling the COVID-19 pandemic” (Lancet 2020)
- Self-management and prevention are key to improved healthcare outcomes, especially highlighted by the COVID-19 pandemic (Rudd 2015, McCormack 2017)
- Health literacy is defined as the ability to obtain, process, and understand medical information (National Academy of Medicine 2004, 2017)
    - Only 12% of US adults are deemed to be ‘proficient’ in health literacy; almost 90 million have low health literacy (Kutner et al. 2006)
    - 60% of Australians are deemed to have low health literacy (https://www.aihw.gov.au/reports/australias-health/australias-health-2018/contents/indicators-of-australias-health/health-literacy)

Use of YouTube for Health Education

- A valuable channel for health education and communication
  - > 100 million+ videos on the diagnosis, treatment, and prevention of a large number of health conditions (Liu et al. 2020)
  - Health promotions , patient education , providing instructions on health procedures
- Criticisms of visual social media use in healthcare (Fernades-Llatas et al. 2017)
  - Videos contain information contradicting reference standards/guidelines
  - Lack a clear and consistent mechanism to retrieve high quality videos for patient education

Research Question

How can we better retrieve medically-relevant and understandable YouTube videos for educating patients and the public on Covid-19?

- Annotating thousands of healthcare related videos requires huge effort
- Automated evaluation of video materials requires a multi-pronged approach
- AHRQ in the US proposed the Patient Education Materials Assessment Tool (PEMAT) to evaluate and compare patient education materials in written, audio and video formats) (Shoemaker et al. 2014)

A video is understandable when

- Consumers of diverse backgrounds and varying levels of health literacy can process and explain key messages

Data Collection Process

- Keyword-based search from three sources:
  - Expert Answers forum on DailyStrength
  - YouTube search suggestions for "COVID-19"
  - Frequently Asked Questions (FAQ) lists of WHO and CDC

- Top 25 videos per keyword from YouTube searches in November 2020
  - Filtering criteria: English language, video duration between 1 to 6 minutes
  - Some videos excluded due to licensing issues

Video Annotation

- Annotated by graduate research associates from consumer perspective, and clinical medicine-trained associates from domain expert perspective
- Criteria for annotation:
  - Medical information (high or low levels)
  - Understandability (rated using PEMAT framework)
  - Overall recommendation (consumer and domain expert perspectives)
- Experts' annotation used to resolve inconsistencies

Video Feature Extraction

- Video metadata retrieval - YouTube Data API
- channel ID (account name), publishing time of the video, title, description, tags, duration, and definition, etc.
- Video level feature retrieval - Google Cloud Video Intelligence API
- shot change detection, object detection, etc.
- Medical terminology extraction - Unified Medical Language System (UMLS)
- Assess text relevance - cosine similarity metrics
    - Metrics compare search keyword to video title, description, and transcription

Video Classification

- Correlation-based feature selection to reduce redundancy in features
- Machine learning algorithms on selected features
  - logistic regression, support vector machine, random forest, XGBoost
- Among 305 videos collected, 40% are recommended by experts
  - 80% of the dataset (194 videos) was used for training
  - 20% (49 videos) was reserved for evaluation
  - Classification performance of the models was assessed on the remaining 61 videos

Evaluation Results

- Logistic regression achieved the highest performance with an accuracy of 83.61%.
- The top five important features identified by the logistic regression classifier:
    - Understandability
    - Medical information expert score
    - Readability score of description
    - Number of shots in the video
    - Cosine similarity between video description and keyword

Classifier	Logistic Regression	SVM	Random Forest	XGBoost
Accuracy	83.61%	77.63%	83.61%	73.68%
Recall	75.86%	66.67%	66.67%	63.33%

Sample Videos

Recommended	Description	Not Recommended	Description
Transmission of COVID-19	Factual and timely guidance for combating COVID-19 provided by two doctors, presented with slides	Battleground states see sharp increases in COVID-19 cases before Election Day	News, minimum medical information presented in the video
COVID-19 Preventing transmission - Environmental cleaning	Educational video from a university with detailed information on Covid-19 prevention	CDC abruptly reverses guidance on COVID-19 airborne transmission	News, low medical information presented in the video

Conclusions

- Evaluating Covid-19 videos on YouTube
  - Assess understandability, medical informativeness, and recommendation suitability
- Potential applications of the pipeline
  - Support clinician decision-making by recommending ranked videos along with instructions for prescriptive usage
  - Assist patients in self-care
- Implications for content creators
  - Use the automated approach to ensure content validity and understandability of the encoded information

Limitations and future research directions

- Addressing variability in video content, representativeness bias and data loss in the videoID collection process
- Exploring additional video features and deep learning models for improved classification performance
- Considering criteria such as content timeliness and accuracy
- Evaluation by clinical experts and consumers through experiments and observational studies

Acknowledgement

We acknowledge support from the NIH/National Library of Medicine grant #R01LM013443 and three graduate students at the Heinz College for assistance with the labeling of the videos.

Thank you!

Questions?