MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

YouTube Video Analytics for Patient Education: An Exploratory Clustering of Obstructive Sleep Apnea Videos

Rema Padman, PhD, FAMIA

Trustees Professor Of Management Science And Healthcare Informatics
The Heinz College of Information Systems & Public Policy
Carnegie Mellon University, Pittsburgh, PA, USA; rpadman@cmu.edu

Co-authors: R. Zhang, MS, J. Shin, MD, K. Schulz, PhD, X. Liu, PhD, A. Susarla, PhD

Funding: NIH/National Library of Medicine #R01LM013443

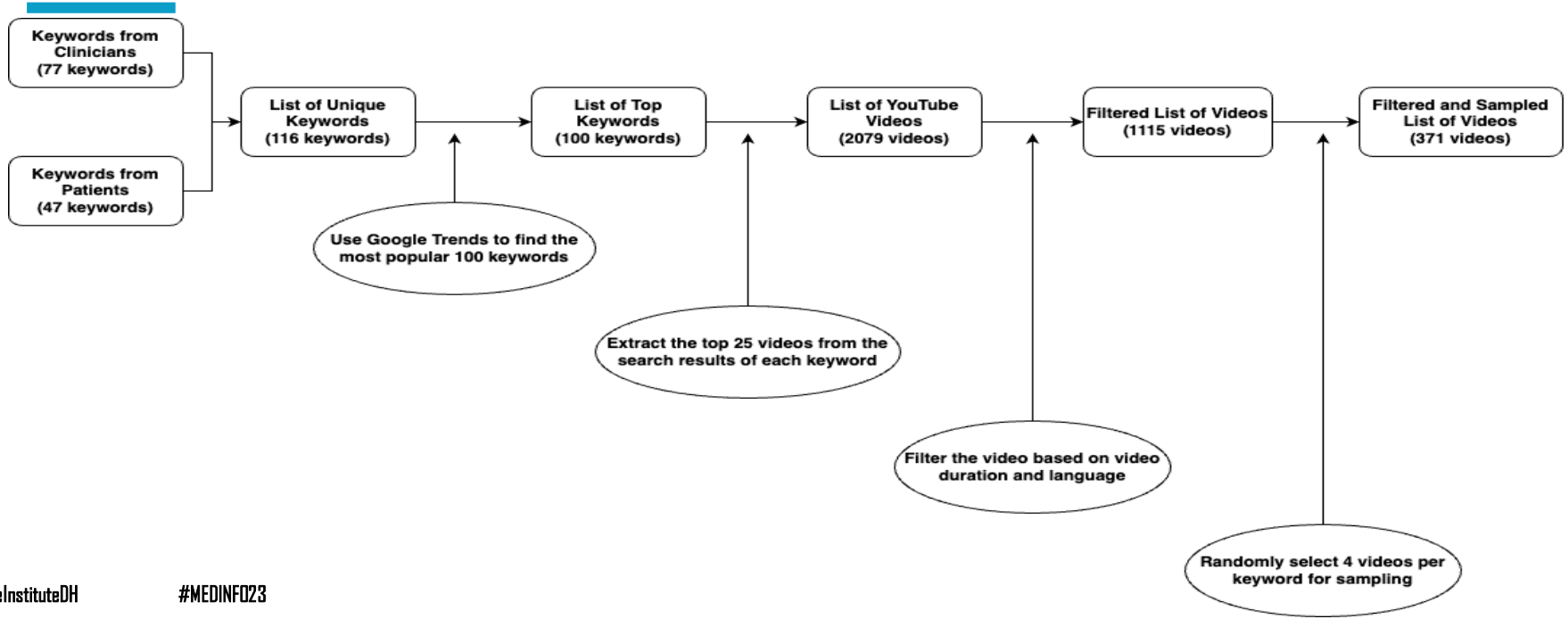
Introduction - YouTube as a Healthcare Information Repository

- YouTube hosts millions of **health-related videos** covering a large variety of health topics from diagnosis and treatments to prevention and self-care (Liu et al. 2020)
- Its **user-generated content** can help improve adherence to clinical guidelines and self-care for managing chronic diseases (Drozd et al. 2019)
- However, **quality issues** in video content may lead to **misinformation and harm**, creating confusion for patients seeking reliable information (Tom et al. 2022)
- Can we curate the large volume of user-generated, visual social media content on the YouTube platform using ML and NLP methods + human annotations to meet multiple criteria – Medical content? Understandability? Actionability? Inclusivity? Accuracy? (Liu et al. 2020,)

Problem - Investigating Obstructive Sleep Apnea (OSA) Videos for Patient Education

- Obstructive Sleep Apnea (OSA) is a sleep-related breathing disorder affecting more than 1 billion adults worldwide (Slowik et al. 2023)
- **We hypothesize** that certain **identifiable features** of OSA-related YouTube videos may reveal distinct patterns that differentiate video clusters
  - These patterns can assist in identifying gaps and challenges for creating more engaging patient education videos
  - Example: Videos of long duration adversely affect patient engagement (Drozd et al. 2018)
- **Our objective** is to explore and identify key features that
  - Allow for a quick evaluation of a YouTube health video before viewing
  - Define distinct video clusters based on these features

Methods – Video Collection Pipeline



Methods - Data Collection and Video Clustering

- We utilized YouTube Data API to collect 8 key features on YouTube videos which we categorize as either intrinsic or extrinsic features
- We performed separate clustering of videos based on these features. Hierarchical clustering was optimal for the differently scaled intrinsic features, while K-means clustering worked well with the more homogeneous extrinsic features

Intrinsic Features

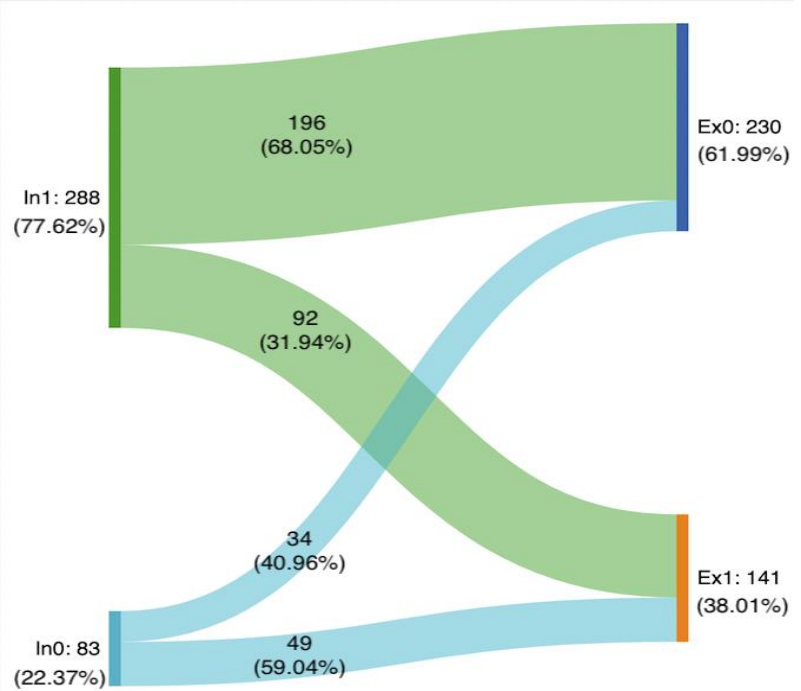
- Duration of the video in seconds
- Number of characters in the video description
- Number of tags associated with the video
- Number of subscribers of the channel that hosts the video
- Text cosine similarity between keyword and video description

Extrinsic Features

- Number of views per day published
- Number of comments the video received per day published
- Number of likes the video received per day published

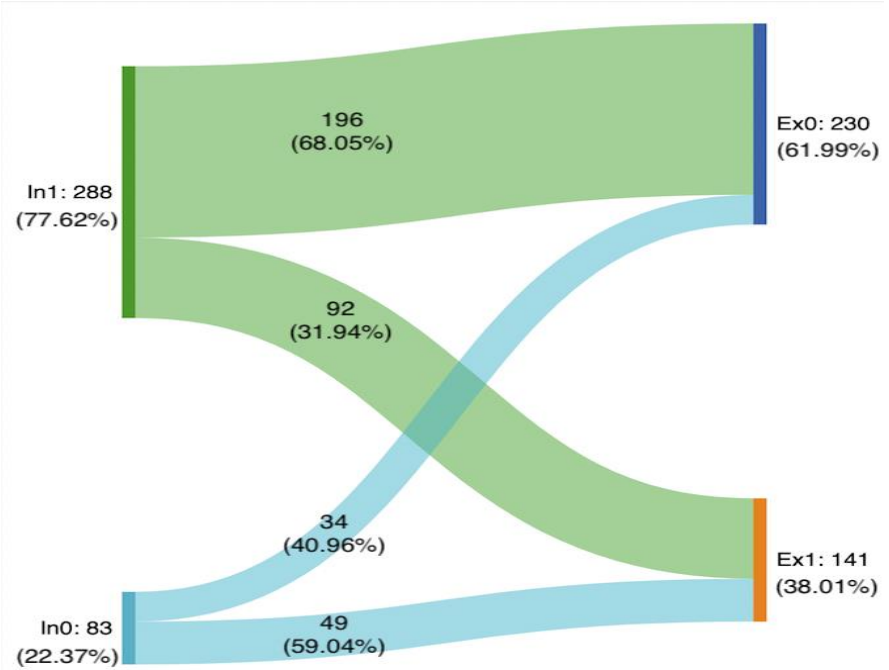
Results: Clustering and Visualization

- A 2-cluster solution was most interpretable for our dataset
- Clustering on intrinsic features resulted in 2 clusters :
  - In1 (78%), more likely to be low-content videos
  - In0 (22%), more likely to be high-content videos
- Clustering on extrinsic features resulted in 2 clusters :
  - Ex0 (62%), more likely to be popular videos
  - Ex1 (38%) , less likely to be popular videos
- Kruskal–Wallis H Test with Bonferroni correction was applied to test for differences across the groups
- All features, except for text cosine similarity between keyword and video description, were found to be statistically different across the clusters



Results: Clustering and Visualization

- Nearly **70%** of the low-content (In1) videos are in the unpopular (Ex0) cluster, and almost **85%** of videos the unpopular (Ex0) cluster originate from the low-content (In1) videos
- High-content (In0) videos are more likely **(59.04% vs 31.94%)** to be included in the popular (Ex1) video cluster than those from the low-content (In1) videos
- Longer duration and better tags/descriptions of high-content videos may attract more viewers compared to those in the low-content cluster



Discussion, Limitations, Future Work, Conclusions

- **For YouTube:** Healthcare videos on YouTube are accessed primarily through searches, rather than browsing
  - Evaluating their intrinsic and extrinsic features can help to filter out low-quality or less valuable content for patient education, potentially enabling more effective algorithmic recommendations
- **For content creators:** Health-related videos, often characterized by fewer tags and shorter descriptions, may not actively promote themselves
  - Recognizing distinguishing intrinsic features could provide key insights for **content creators** to improve video popularity
- **Limitations:** The limited number of videos available for analysis was a challenge in this study
- **Future work:** Expand video selection, incorporate content analysis techniques such as sentiment analysis, develop a recommendation classifier to identify high-quality videos for patient education and evaluate their impact on learning and health outcomes
- **Conclusions:** A feature-based approach to clustering of health-related YouTube videos on OSA using statistical and machine learning methods allows for rapid assessment of important clusters and their features for further analysis, and subsequent classification into high and low quality videos

Acknowledgement

- We acknowledge support from the NIH/National Library of Medicine grant #R01LM013443.

Thank you! Questions?