*MEDINFO 2023 — The Future Is Accessible*
*J. Bichel-Findlay et al. (Eds.)*


*doi:10.3233/SHTI231167*

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

Ruoyu ZHANG[a], Jennifer SHIN[b], Kristine SCHULZ[b], Xiao LIU[c], Anjana SUSARLA[d] and Rema PADMAN[a,1]
[a] *Carnegie Mellon University, Pittsburgh, Pennsylvania, USA*
[b] *Brigham and Women's Hospital, Boston, USA*
[c] *Arizona State University, Tempe, USA*
[d] *Michigan State University, Lansing, USA*

**Abstract.** With the growing popularity of content-sharing platforms, patients are increasingly using the Internet as a critical source of health information. As one of the most popular video-sharing sites, YouTube provides easy access to health information seekers, but it is difficult and time-consuming to identify and retrieve high-quality videos that may serve as engaging patient education materials. This paper reports on an exploratory analysis of 317 YouTube videos on Obstructive Sleep Apnea (OSA) to better understand some key features of the videos and the relationships between them to facilitate subsequent video classification and recommendation. Features intrinsic to a video, such as video duration, and extrinsic, such as the number of views, are analyzed using unsupervised clustering methods and the Sankey diagram to discover the relationship between the clusters and their significance across different clusters, providing promising insights for the assessment of video quality.

**Keywords.** Obstructive sleep apnea, youtube videos, clustering methods

## 1. Introduction

The easy availability of a vast repository of computable biomedical and human-centered, user-generated, health information on the YouTube social media platform presents an unprecedented opportunity to investigate how social media can be an engaging channel to inform and communicate healthcare information to patients and facilitate patient-centric health promotion and literacy improvement. YouTube hosts millions of healthcare related videos on the pathogenesis, diagnosis, treatment, and prevention of a variety of medical conditions [1]. This plethora of user-generated content may be mobilized to improve adherence to clinical guidelines and self-care required for management of chronic diseases [1]. However, the widely differing video content quality raises concerns in the context of patient education [1, 2]. Patients might find themselves troubled by misinformation and disinformation when they use health-related keywords on YouTube and need better tools to filter the videos in the search results. We hypothesize that video features may capture different aspects of information about

[1]Corresponding Author: Rema Padman, email: rpadman@cmu.edu.

health-related videos and there may be distinct patterns that differentiate the video clusters. Profiling these patterns may help content creators and end users to identify gaps and challenges in creating videos for the specific purpose of patient education. For example, if the majority of videos available on a health topic of interest to patients have a long duration, studies have shown that patient engagement with the video will be adversely affected [1], opening an opportunity to produce new videos that are of shorter duration. Our goal is to discover distinct clusters of videos defined by different key features of healthcare-related YouTube videos and explore what features can be used to quickly evaluate a YouTube video before even viewing it.

## 2. Methods

### *2.1. Video Collection*

Our dataset included 371 YouTube videos on the topic of Obstructive Sleep Apnea (OSA). These videos were collected using two lists of search keywords related to OSA, created by co-authors JS and KS, that clinicians and patients, respectively, would use to search for patient education videos on the YouTube platform. The clinician list included 77 keywords, such as Daytime sleepiness, Snoring, and Wake gasping. The patient list had 47 keywords, such as Obstructive sleep apnea, Airway pressure, and Acromegaly. We combined the two lists into one with 116 unique keywords about OSA, its definition, treatments, and so on, then applied Google Trends, a website that analyzes popularity of top search queries in Google Search, to find the top 100 out of 116 keywords in popularity to focus video collection on most common search queries.

For each selected keyword, 4 videos were randomly selected for downloading from the first page of search results on the YouTube website to build a representative dataset for this exploratory study. Since the same video may appear in search results of different keywords ("cause of OSA" and "OSA" may give similar search results), duplicates were removed and filtered further for short videos, between 1 to 6 minutes long [1], and in the English language, resulted in 371 videos for our analysis.

### *2.2. Metadata Collection and Pre-processing*

YouTube Data API is a powerful tool provided by Google to help developers and researchers collect key video features which we categorize as intrinsic and extrinsic features. Intrinsic features can be considered as properties of the video such as its duration, language, and so on that remain static over time. Extrinsic features are exogenous to the video such as the number of views, likes and comments that change over time once the video is published. Five intrinsic and three extrinsic features that are commonly available for every video were downloaded using the API for our metadata analysis. Table 1 shows the feature name, its corresponding description, and its type.

We derived some features from the downloaded metadata for better interpretability. For example, the number of characters in video description measured how detailed it was; cosine similarity, ranging from 0 to 1, was computed to capture the similarity between users' search keywords and video description; the total views, comments, and likes for each video were normalized by dividing by its published time and used as the three extrinsic features. Some features such as likes and comments are highly right-skewed and of larger scale, hence we apply log transformation [4] on channel subscribers,

views, likes, and comments to decrease variability and approximately conform to normality.

**Table 1.** Summary of Intrinsic and Extrinsic Features.

| Feature Name | Description | Intrinsic/ Extrinsic |
|---|---|---|
| Duration | Duration of the video in seconds | Intrinsic |
| Description Length | Number of characters in the video description | Intrinsic |
| Number of Tags | Number of tags associated with the video | Intrinsic |
| Channel Subscribers | Number of subscribers of the channel that hosts the video | Intrinsic |
| Cosine Similarity | Text cosine similarity between keyword and video description | Intrinsic |
| Views | Number of views per day published | Extrinsic |
| Comments | Number of comments the video received per day published | Extrinsic |
| Likes | Number of likes the video received per day published | Extrinsic |

### *2.3. Clustering Methods*

We cluster the videos separately on intrinsic and extrinsic features and then map them across the two categories to visualize potential relationships between them that can subsequently be rigorously analyzed using statistical and machine learning methods. We apply two popular clustering methods for our task - hierarchical clustering and K-means clustering. Features in the extrinsic dataset are more homogenous and less skewed than features in the intrinsic dataset, allowing K-means clustering to work well. Intrinsic features differ in scale, hence hierarchical clustering is a better approach for this dataset. An experimental evaluation of 2, 3 and 4 clusters with both datasets determined that the combination of 2 clusters is optimal and more interpretable for the OSA video dataset.

## 3. Results

### *3.1. Results of Clustering*

Clustering on intrinsic features resulted in 2 clusters with 288 videos (~78%) labeled In1, and 83 videos (~22%) labeled In0, respectively, whereas clustering on extrinsic features produced 2 clusters with 230 (62%) labeled Ex0 and 141 (38%) videos labeled Ex1, respectively. Thus, after clustering, each video is associated with both an intrinsic cluster and an extrinsic cluster. Table 2 provides summary statistics of the features across all the clusters. On average, In1 videos have shorter duration, shorter description length, fewer tags, and fewer channel subscribers in comparison to In0 videos, indicating higher likelihood of being low-content videos. Similarly, Ex0 videos have fewer views, comments, and likes than Ex1 videos which are higher on the popularity measurements.

To better understand the relationship between intrinsic and extrinsic features, we use a Sankey diagram to display the flows between the clusters. A Sankey diagram is a visualization that depicts flows from the two ‘sources’ of intrinsic clusters to the two ‘targets’ of extrinsic clusters, as shown in Figure 1. We observe that almost 70% of the In1 videos are in the unpopular Ex0 video cluster and almost 85% of the videos in Ex0 come from In1. The more in-depth In0 videos are more likely (59.04% versus 31.94%) to be included in the popular cluster of Ex1 videos than members from the In1 videos. It is plausible that longer duration and better tags and descriptions invite more viewers to access the video than those in In1 cluster.

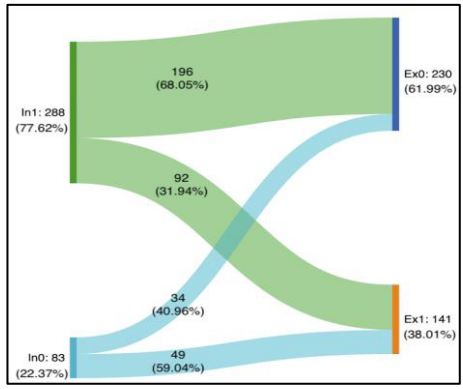


**Figure 1.** Sankey Diagram of the clustering results.

**Table 2.** Clustering Results on Extrinsic and Intrinsic Features.

| **Feature Name** | **Total** | **(In, Ex) = (0, 0)** | **(In, Ex) = (0, 1)** | **(In, Ex) = (1, 0)** | **(In, Ex) = (1, 1)** |
|---|---|---|---|---|---|
| | **N = 371** | **N = 34** | **N =49** | **N =196** | **N =92** |
| Duration | 199.93±85.5 | 206.21±92.2 | 236.59±69.5 | 180.6±86.12 | 219.25±79.3 |
| Cosine Similarity | 0.07±0.09 | 0.07±0.04 | 0.06±0.04 | 0.07±0.10 | 0.08±0.10 |
| Description Length | 945.2±1009 | 2586.3±945 | 2474.7±810.5 | 392.2±356.7 | 702.2±460.2 |
| Number of Tags | 12.47±10.26 | 16.12±11.1 | 21.55±10.06 | 9.45±8.62 | 12.72±10.07 |
| Channel Subscribers | 10.86±3.32 | 10.83±2.68 | 13.56±2.13 | 9.45±3.16 | 12.41±2.78 |
| Views | 10.39±2.89 | 8.71±1.76 | 13.24±1.94 | 8.77±2.16 | 12.94±1.80 |
| Comments | 3.01±2.65 | 1.91±1.26 | 5.77±2.07 | 1.30±1.40 | 5.58±1.98 |
| Likes | 5.73±3.00 | 4.35±1.74 | 9.11±2.04 | 3.85±1.93 | 8.44±1.84 |

### *3.2 Significant Features across Clusters*

Since each video can be in two distinct clusters based on extrinsic and intrinsic features, there are four groups of interest- In1 (likely low content), In0 (likely high content), Ex1 (likely unpopular) and Ex0 (likely popular) - to examine for critical features that define them. Therefore, we apply the Kruskal–Wallis H Test with Bonferroni correction to test whether the average value of the features is statistically different across different groups. The results are reported in Table 3. All features other than cosine similarity can be seen to be different across the clusters.

**Table 3.** Significant features using Kruskal-Wallis H Test.

| **Feature Name** | **p value** | **Feature Name** | **p value** |
|---|---|---|---|
| Duration | <0.0001 | Channel Subscribers | <0.0001 |
| Cosine Similarity | 0.3955 | Views | <0.0001 |
| Description Length | <0.0001 | Comments | <0.0001 |
| Number of Tags | <0.0001 | Likes | <0.0001 |

## 4. Discussion

Unlike music videos or news, people generally access healthcare videos on YouTube by searching instead of browsing. Consumers may watch the music video of a song they

have never heard, but most people will not watch a video about a disease that the individual or their family members do not have. Healthcare videos do not generally appear in your recommendation flow and they are comparatively unpopular based on the number of views or likes. As more patients, their caregivers and the public seek health information online, it is important to discover what kind of videos would be returned by a search keyword and whether a quick review of its intrinsic and extrinsic features may provide sufficient information to filter out videos that are likely to be of low quality or limited value for patient education [3]. The resultant video set could potentially be delivered as an algorithmic recommendation.

Health-related videos generally do not need to promote themselves. As a result, they have fewer tags and shorter descriptions, which define the In1 videos. This may provide valuable insights to health content creators to pay attention to tags and descriptions as a signal mechanism to create more popular videos. Besides discovering the underlying patterns, we identify important features that distinguish the clusters and differentiate them. The result indicates that duration, length of description, number of tags, channel subscribers, views, comments, and likes vary with the cluster. Ongoing analyses will quantify the feature importance using statistical models.

A major limitation in building more rigorous models in this study is the limited volume of videos available for analysis. Future work will expand our selection of videos and apply content analysis to increase the number of relevant features such as analysis of video transcripts for eliciting sentiment and extracting medical information from the description using Named Entity Recognition (NER) methods [1]. These features will further be used to build a recommendation classifier for identifying high-quality videos for patient education.

## 5. Conclusions

In this paper, we apply a feature-based approach to cluster and evaluate health-related YouTube videos on OSA, providing a quick assessment of important clusters and their features for further analysis, and subsequent classification into high and low quality videos using statistical and machine learning methods.

## Acknowledgement

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

## References

[1] Liu X, Zhang B, Susarla A, Padman R. Go to youtube and call me in the morning: use of social media for chronic conditions. MIS Q. 2020 Mar; 44(1):257-83, doi: 10.25300/MISQ/2020/15107.

[2] Drozd B, Couvillon E, Suarez A. Medical youtube videos and methods of evaluation: literature review. JMIR Med Educ. 2018 Feb;4(1):e3, doi: 10.2196/mededu.8527.

[3] Tom K, Phang PT. Effectiveness of the video medium to supplement preoperative patient education: a systematic review of the literature. Patient Educ Couns. 2022 Jul;105(7):1878-87, doi: 10.1016/j.pec.2022.01.013.

[4] Feng C, Wang H, Lu N, Chen T, He H, Lu Y, Tu XM. Log-transformation and its implications for data analysis. Shanghai Arch Psychiatry. 2014 Apr;26(2):105-9, doi: 10.3969/j.issn.1002-0829.2014.02.009.