MACHINE LEARNING APPROACHES FOR ACCESSIBLE PUBLIC HEALTH INFORMATION ABOUT COVID-19

Anonymous authors
Paper under double-blind review

1 BACKGROUND

Health literacy is well recognized as a challenge for public health, with many adults lacking the requisite skills to engage successfully in the management of their healthcare. Most health education materials remain too complicated for patients to comprehend (Nutbeam, 1998), which is amplified when considering that 80% of Americans search for health information on the internet and only 12% have proficient health literacy to correctly interpret and use it (Nutbeam, 1998; Sørensen et al., 2012). YouTube provides metadata about videos in a standard and semi-structured format, which can be processed by computers. However, existing machine learning approaches do not parse the semantic meaning of video content or their understandability, especially in the medical domain. In this abstract, we describe research in progress to address this challenge by identifying a limited number of short duration, publicly available videos on the YouTube social media platform that clinicians and public health experts can recommend for disseminating medically relevant and understandable multi-media rich information on COVID-19 and educating the public on the pandemic. Building on our prior research, we aim to (1) assemble a corpus of COVID-19-related YouTube videos that have exploded on the social media platform in recent months, and (2) develop an augmented intelligence-based approach to analyze them using deep learning, image analytics and natural language processing (NLP) methods with evaluative inputs from an infectious disease expert.The overall goal is to retrieve the top few videos for public dissemination that meet the understandability and medical content criteria.

Most existing work in health communication and patient education have evaluated patient education videos manually (add some references), which is not scalable or cost efficient. We apply the Patient Education Materials Assessment Tool (PEMAT) (Shoemaker et al., 2014), developed by the Agency for Healthcare Research and Quality (AHRQ), to examine whether videos are considered understandable from a patient education perspective. We will collect 50 commonly used COVID-19 search keywords from Google Trends, journal articles, health magazines and from our infectious disease expert/consultant and apply them to retrieve a corpus of YouTube videos and their metadata. Two graduate students with clinical expertise will watch 25% of the video collection, assess the amount of COVID-19 related medical information in the videos and their understandability and label them positively/high or negatively/low based on these two dimensions of interest. We will use Cohen's kappa statistic to measure inter-rater reliability (Vishnevetsky et al., 2018). The infectious disease expert/consultant will review any disagreements and make final judgements. We will develop a multi-view co-training approach (Blum & Mitchell, 1998) that will analyze YouTube videos using their metadata and features and assess the understandability of the video according to PEMAT guidelines. In case the multi-view model yields inconsistent results for videos, the infectious disease expert/consultant will review them and provide a consensus opinion.

We will extend a Bidirectional Long Short Term Memory (BLSTM) model to extract medical terms from the user-generated video descriptions at the sentence level that was developed in prior work (Liu et al., 2019). We will classify the encoded medical information in videos into high and low information groups. We will then determine whether a video should be recommended to the public, patients and clinicians by generating a video quality index according to both understandability and medical information. Two graduate students with clinical expertise will review the videos and label a subset of them that meet both criteria. A logistic regression model will be built based on this assessment. The classifier will then be used to automatically recommend the top videos for COVID-19 related patient education, public health awareness and health communication. We will conduct a computational evaluation of our proposed methods based on accuracy, precision, recall, and f-measure. We will also compare our models with feature-based classification methods and medical expert's recommendations.

By enabling a generalizable, automated approach to retrieve, label and recommend COVID-19 related videos with high medical information content and understandable by laypersons, and providing a scalable technology prototype that embeds AI and machine learning and is validated by public health guidelines, our research outputs can be used by decision makers to develop best practices to promote adherence to advice from public health experts. The video recommendations may also be disseminated directly to the public to inform and educate. This approach has the potential to improve health literacy disparities for a challenging societal problem.

2 VIDEO LABELING

YouTube videos collected for this study should be relevant to patient education and represent what YouTube searches returned to patients/consumers. We will collect 50 commonly used COVID-19 search keywords as mentioned earlier and apply them to retrieve a corpus of videos and their metadata. We will collect the top 100 videos for each search term and store the ranking of returned videos and their metadata in a database for further analysis. Using YouTube Data API, we will gather information on channel ID (account name), publish time of video, video title, video description, video tags, video duration, video definition, video caption availability, video rating, view count, like count, dislike count, and comment count. In total, we will work with almost 5,000 unique videos using the 50 search terms, after removing duplicates. We will also collect the video content (i.e., the actual video files) from the platform using an open source tool YouTubeDL. This corpus of COVID-19 related YouTube videos will comprise the data for the proposed study.

The gaps in available health information materials to meet the health literacy level-based needs of consumers has prompted the Agency for Healthcare Research and Quality (AHRQ) to develop guidelines for patient educational materials called Patient Education Materials Assessment Tool (PEMAT) (https://www.ahrq.gov/ncepcr/tools/self-mgmt/pemat.html). PEMAT assesses the domains of understandability of healthcare information, i.e., when consumers of diverse backgrounds and varying levels of health literacy can process and explain key messages (Shoemaker et al., 2014). We will assess 10% of the entire collection for understandability according to the PEMAT guidelines. Some basic requirements for selecting videos for labeling include (1) video length < 6 minutes (Liu et al., 2019); and (2) video includes a narrative in English to enable video data analysis for feature extraction (described in the next section). We will recruit both domain experts and consumers to label the videos using the PEMAT criteria. Sample size calculation indicates that less than 500 videos are needed to achieve high inter-rater reliability ($K > 0.80$) (Vance et al., 2009). However, since machine learning methods require large number of videos for training and testing, our labeling corpus will include more videos than the minimum required. We will assess inter-rater reliability using Cohen's kappa (Vishnevetsky et al., 2018). The large number of videos in the labeled corpus will allow us to select only the subset of videos where there is perfect match (including both high and low understandability) in order to build the co-training model.

3 VIDEO ANALYSIS METHODS FOR FEATURE EXTRACTION

We will develop two classifiers from two sufficient and conditionally independent views (i.e., video metadata and video content) and assess video understandability. The Google Cloud Video Intelligence platform (https://cloud.google.com/video-intelligence/docs/) supports object detection, scene change detection, video transcription, facial expression detection and optical character detection. It has pre-trained ML models that automatically recognize a vast number of objects, places, and actions in stored or streaming videos. Object detection can help us to extract labels of objects in the videos at frame level and shot level. The platform can detect scene changes within videos which help us examine whether videos are breaking into small sections. Speech transcription can extract narratives. Optical character recognition can detect and extract text, tables, or illustration in videos. Tables 1 and 2 summarize the features we will extract from video metadata and video content views, methods to derive the measures, and the PEMAT criteria they meet.

Each video's metadata contains video title, description, and tags, which are submitted by the content creator. A video with good understandability to patients uses common daily language. The Auto-

Table 1: Features for Video Understandability Classification from Video Metadata View

Feature Name	Method	PEMAT Criterion
Has title	Metadata collection	The material makes its purpose evident
Has description	Metadata collection	
Has tags	Metadata collection	
Description Readability	Readability analysis	The material uses common every-day language
Active word count	Syntactic analysis	The material uses active voice
Summary word count	Semantic analysis	The material provides a summary
Transition word count	Semantic analysis	The material presents information in a logical sequence
Video duration	Metadata collection	Medical information encoded in the video (Liu et al., 2019)
Description word count	Metadata collection	
Sentence count	Metadata collection	
Description unique words	Metadata collection	
Description medical term count	Medical entity recognition	

mated Readability Index (ARI) is a test designed to assess the understandability of a text and has been widely adopted to evaluate health information and resources online (Yang et al., 2015). We will perform syntactic analysis to extract the number of verbs in active voice with Part-of-speech tagging. Semantic analysis focuses on evaluating the relations of sentences (e.g., in proper logical sequence, with transitions, and summaries). The use of transition words and phrases can improve the logical connections in writing and speech. Transition words and phrases can be grouped into categories such as causation, chronology, combinations, contrast, example, clarification, summary, and more. We will compute the number of summary words to determine whether the material provides a summary and apply text preprocessing techniques to identify the total number of words, sentences, and unique words from the video description. The video content view generates features using video

Table 2: Features for Video Understandability Classification from Video Content View

Feature Name	Method	PEMAT Criterion
Narrative readability	Readability analysis	The material uses common every-day language
Active word count	Syntactic analysis	The material uses active voice
Summary word count	Semantic analysis	The material provides a summary
Transition word count	Semantic analysis	The material presents information in a logical sequence.
Video transcription confidence	Auto transcription	The material allows users to hear the words clearly
Text detection confidence	Optical character recognition	The text on screen is easy to read
Shot count	Shot detection	The material breaks or "chunks" information into short sections.
Transcript word count	Auto transcription	Medical information encoded in the video (Liu et al., 2019)
Transcript unique word	Auto transcription	
Transcript sentence count	Auto transcription	
Transcript medical term	Medical entity recognition	
Video object	Object detection	

narratives, shots and associated confidence scores. We will use Google Video Analytics tools to detect shot changes, transcribe the videos, and recognize optical characters in the videos. The number of scenes in the video is an indicator if the video has short sections. The narrative readability score examines whether the material uses common everyday language. Part-of-speech tagging is used to extract verbs in active voice in the transcript. The number of transition words and summary words is identified according to the transition word list. The video transcription confidence score is a proxy of whether the users can hear the words in narratives clearly. Videos are often broken into different chunks by scenes. The confidence score of text detection indicates the readability of the text on the screen.

We will conduct a computational evaluation of our proposed co-training approach for accuracy, efficiency, parsimony, and interpretability based on precision, recall, and f-measure when classifying videos as high or low understandability videos. We will also compare our model results with those from three feature-based methods (e.g., logistic regression, Support Vector Machines, and Random Forest) using labeled data alone to test the efficacy of our co-training method and with alternative semi-supervised learning methods such as Gaussian mixture models, semi-supervised SVM, and self-training.

4 Algorithm Comparison with Medical Expert Recommendation

To evaluate the significance of video content and understandability in medical experts' decision to recommend a YouTube video for patient education, we will retrieve results from YouTube for the top 10% search keywords from the original keyword set used for video data collection. Our medical experts will review the top 10 videos according to our re-ranked results for each query and report whether they would recommend the videos to patients. Precision at K is a common information retrieval measure used in modern (web-scale) information retrieval systems (Zheng et al., 2017). Precision at K (P@K) assesses how many of the top K results are relevant (e.g., P@10 or "Precision at 10" corresponds to the number of relevant results among the top 10 documents). We measure the average precision at K, with K = 1,. . . , 10 for the top 20 queries.

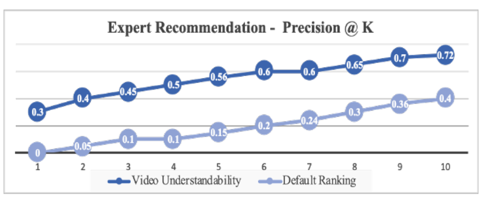

Figure 1: Illustrative Comparison of Video Understandability Ranking with Expert Recommendation

Figure 1 shows an illustrative comparison of the significance of video understandability ranking, where the expert recommends 30% of top 1 videos in understandability. None of the top 1 videos according to the default YouTube ranking is recommended; 72% of top 10 videos in understandability are recommended by experts while only 40% of top 10 videos with the default ranking on YouTube are recommended.

5 Limitations

We retrieve videos from the YouTube platform whose algorithms and methods of video retrieval are proprietary knowledge. Hence, biases may be introduced in the training data from retrieving these videos. We will mitigate this limitation to an extent by conducting multiple levels of human/expert evaluations. In the future, we intend to build de-biasing methods into our algorithms and platform to address this concern.

6 Summary

Our design and implementation of a scalable, automated, and generalizable approach will break through the inability of expert-based methods to assess understandability and encoded medical content in video data at scale. This approach is developed within the context of patient educational video design that can be generalized to content design for different purposes. Overall, the capabilities developed and introduced in this work have widespread usefulness given the ubiquity of social media and multimedia content. Applications include product classification in e-commerce where e-commerce websites typically employ editors and crowdsourcing platforms such as Amazon Mechanical Turk (AMT) to classify products. Integrating text and images from customer reviews into automated machine learning based classification approaches using co-training algorithms may also improve classification accuracy and enable customers to obtain benefits of content and context-based recommendations.

REFERENCES

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100, 1998.

Xiao Liu, Bin Zhang, Anjana Susarla, and Rema Padman. Go to youtube and call me in the morning: use of social media for chronic conditions. *Forthcoming, MIS Quarterly*, 2019.

Don Nutbeam. Health promotion glossary. *Health promotion international*, 13(4):349–364, 1998.

Sarah J Shoemaker, Michael S Wolf, and Cindy Brach. Development of the patient education materials assessment tool (pemat): a new measure of understandability and actionability for print and audiovisual patient information. *Patient education and counseling*, 96(3):395–403, 2014.

Kristine Sørensen, Stephan Van den Broucke, James Fullam, Gerardine Doyle, Jürgen Pelikan, Zofia Slonska, and Helmut Brand. Health literacy and public health: a systematic review and integration of definitions and models. *BMC public health*, 12(1):1–13, 2012.

Karl Vance, William Howe, and Robert P Dellavalle. Social internet sites as a source of public health information. *Dermatologic clinics*, 27(2):133–136, 2009.

Julia Vishnevetsky, Chasity Burrows Walters, and Kay See Tan. Interrater reliability of the patient education materials assessment tool (pemat). *Patient education and counseling*, 101(3):490–496, 2018.

Shihao Yang, Mauricio Santillana, and Samuel C Kou. Accurate estimation of influenza epidemics using google search data via argo. *Proceedings of the National Academy of Sciences*, 112(47): 14473–14478, 2015.

Nan-ning Zheng, Zi-yi Liu, Peng-ju Ren, Yong-qiang Ma, Shi-tao Chen, Si-yu Yu, Jian-ru Xue, Ba-dong Chen, and Fei-yue Wang. Hybrid-augmented intelligence: collaboration and cognition. *Frontiers of Information Technology & Electronic Engineering*, 18(2):153–179, 2017.