Identifying Features Impacting Healthcare Videos on Youtube Using Face Recognition Methods and BERT Model

By

Jahnavi Chadalavada
Computer Science Department
University of Central Missouri Lees
Summit, MO, USA

Motivation

- Recent research in Information Systems (IS) suggests that there is an increasing usage towards videos to administer self – health.Studies suggest that one in three US adults use the Internet to diagnose or learn about a health concern. However, such access to health information online could exacerbate the disparities in health information availability and use.
- Even though social media offers an excellent conduit for encouraging patients to increase their knowledge on health literacy, there is a significant challenge in producing fair and transparent recommendations that address the needs of every user.
- we need a fair and bias-minimizing approach that ensures recommendations are not skewed against a particular demographic group or set of ideas.
- Hence there is need to understand which features has more impact on views of healthcare videos.

Objectives

This project is an empirical valuation, considering YouTube videos as the primary dataset, and trying to identify features influencing video engagement.

- Diabetes-related videos are collected from YouTube and their metadata using YouTube Data API.Around 350 videos are collected using most relevant search words related to diabetes.
- YouTube Videos are Converted into frames and using OpenCv.
- Extract gender,age and race information using DeepFace framework.
- Retrieve medical information using BERTopic.
- Run ML algorithms to identify which feature has more influence on view count of videos.

Related work

YouTube for Patient Education: A Deep Learning Approach for Understanding Medical Knowledge from User-Generated Videos [1].

Identifying content unaware features influencing popularity of videos on YouTube: A study based on seven regions [2].

Hyperextended lightface: A facial attribute analysis framework [6].

Problem Statement

To produce fair and transparent recommendations of healthcare videos.we need to understand the biases associated with demographic characters of actors/presenters and medical information in the healthcare videos on youtube.

Hence the need to understand which features impact views of healthcare videos is very high.

Proposed Solution

Project structure has three main steps as below.

Data Collection

The primary source for data is YouTube.YouTube API is used to extract data from the YouTube.Videos are downloaded using youtube video downloaders.

Data preprocessing

- To extract facial attributes first we need to convert videos into frames.OpenCV (Open Source Computer Vision Library) library is used to convert videos into frames.

- To extract facial attributes DeepFace is used.DeepFace is a popular open-source face recognition and facial attribute analysis library that is built on top of Keras,TensorFlow, and OpenCV.

- Perform Topic Modelling to extract medical information from video subtitles using BerTopic [4].Bertopic is a topic modeling algorithm that uses a pre-trained BERT (Bidirectional Encoder Representations from Transformers) language model to encode text data and cluster similar documents into topics.

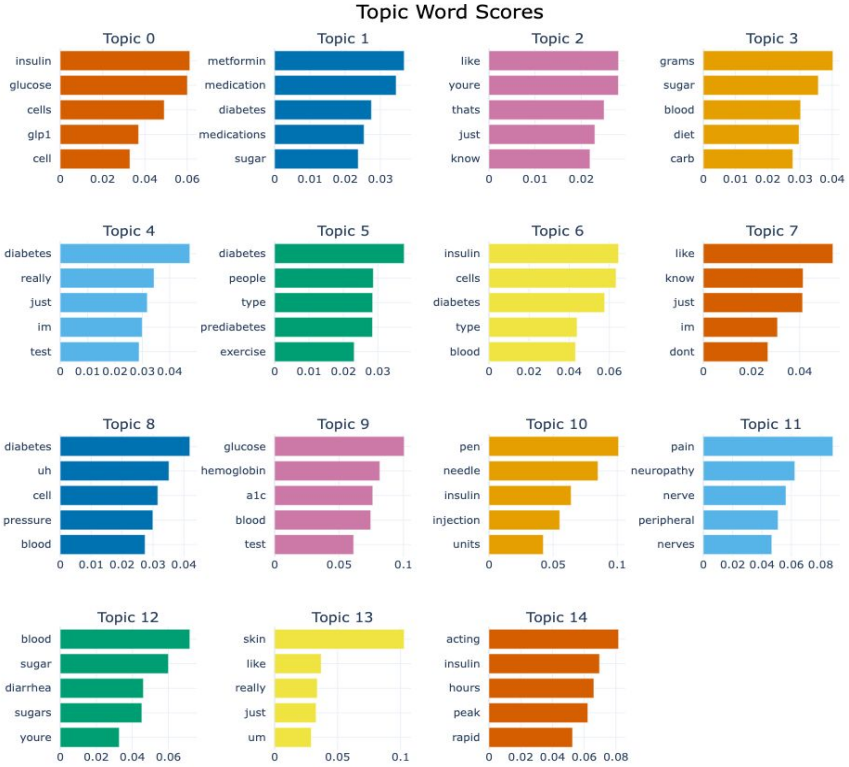
Data analysing using Machine learning techniques.

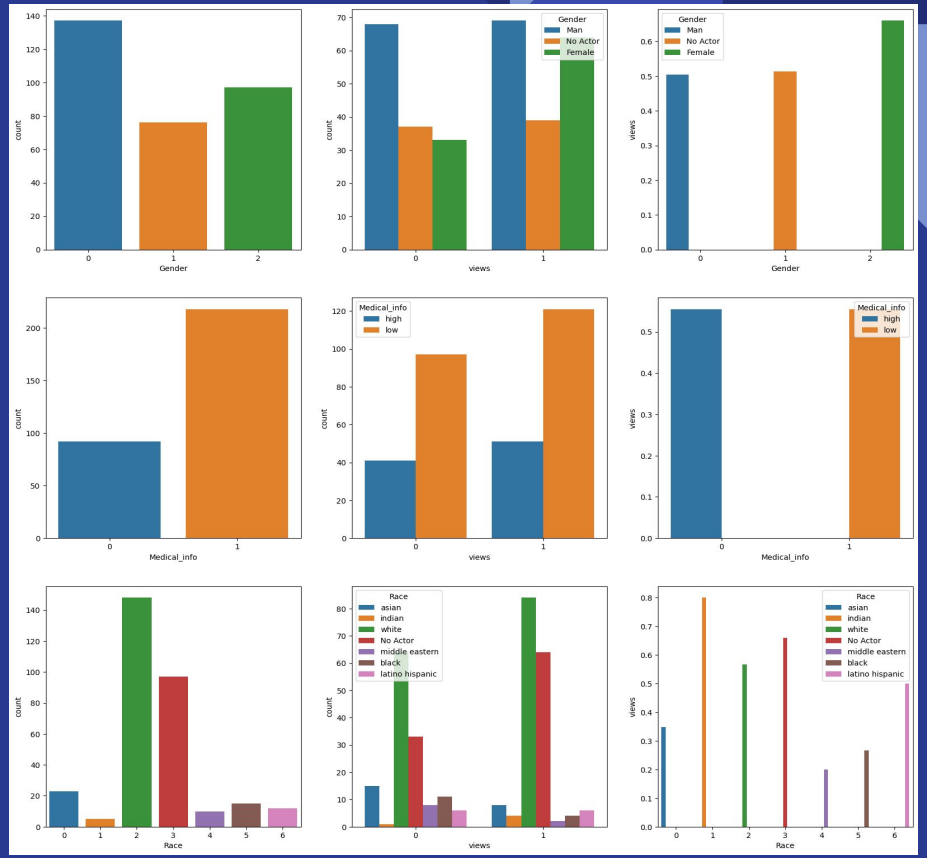
- Knowing which features are important in a machine learning model is essential for understanding how the model is making its predictions and for optimizing the model's performance.

- One of the simplest methods to assess the importance of features is by analyzing the coefficients of the model.

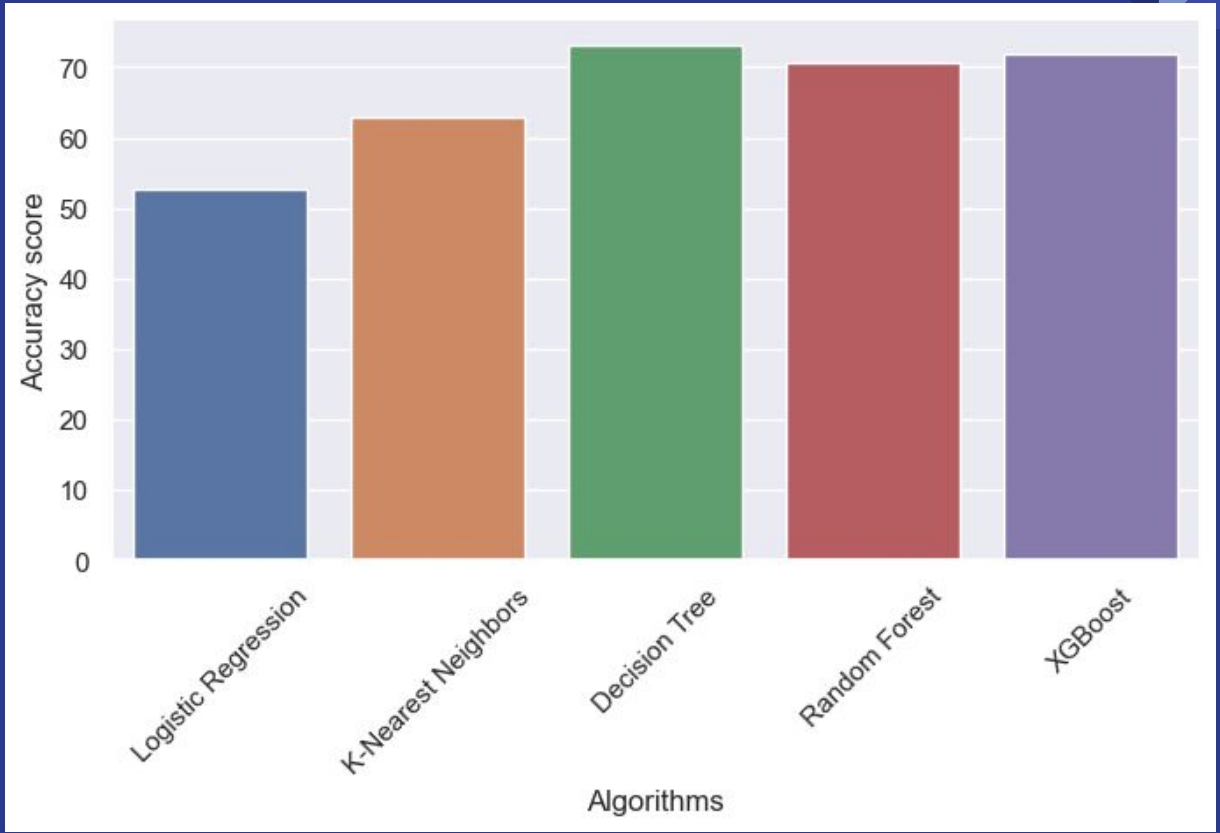
- Many machine learning algorithms, such as logistic regression,decision trees, random forests, and gradient boosting models, provide feature importance scores that indicate how much each feature contributes to the model's performance. These scores can be used to identify the most important features in the data.For example, in scikit-learn, you can use the feature importances attribute of a trained tree-based model to retrieve the feature importance scores.

Results

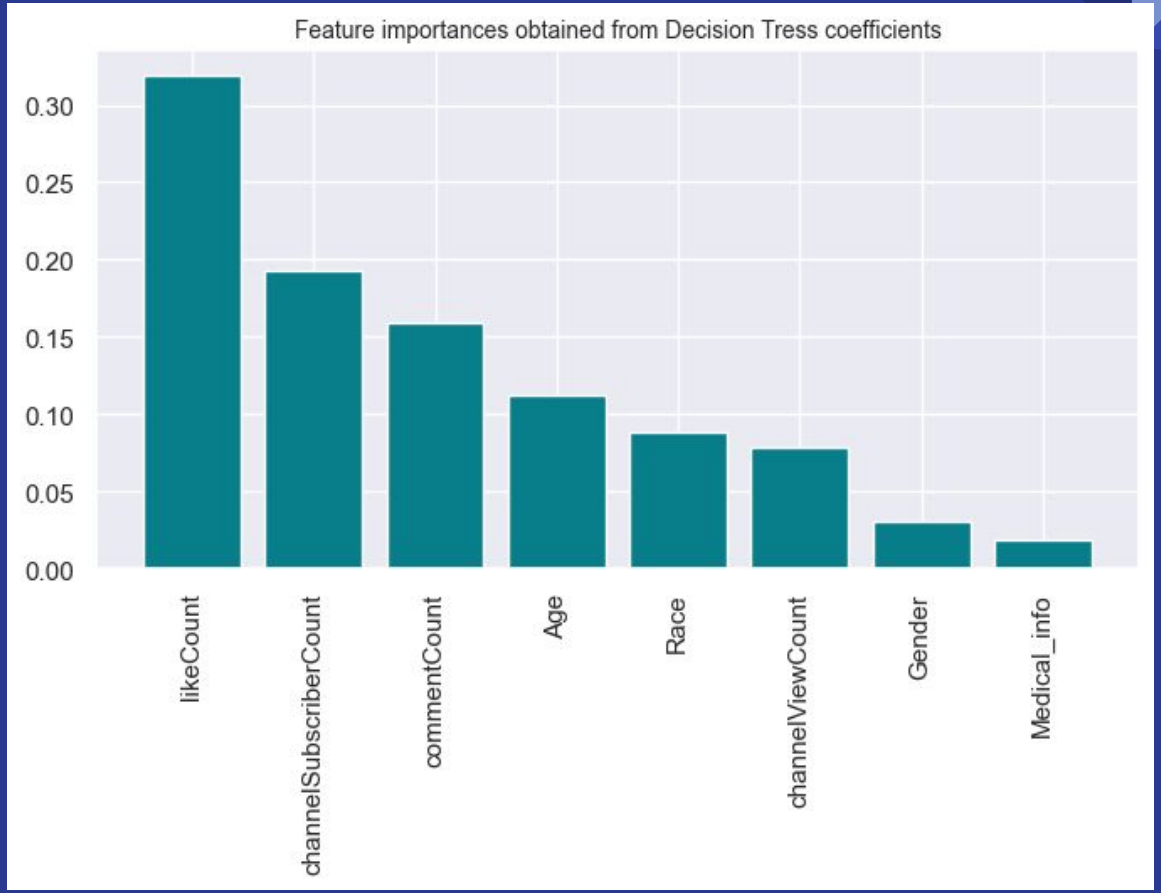

Topic Word Scores
Topic 0
insulin
glucose
cells
glp1
cell
Topic 1
metformin
medication
diabetes
medications
sugar
Topic 2
like
youre
thats
just
know
Topic 3
grams
sugar
blood
diet
carb
Topic 4
diabetes
really
just
im
test
Topic 5
diabetes
people
type
prediabetes
exercise
Topic 6
insulin
cells
diabetes
type
blood
Topic 7
like
know
just
im
dont
Topic 8
diabetes
uh
cell
pressure
blood
Topic 9
glucose
hemoglobin
a1c
blood
test
Topic 10
pen
needle
insulin
injection
units
Topic 11
pain
neuropathy
nerve
peripheral
nerves
Topic 12
blood
sugar
diarrhea
sugars
youre
Topic 13
skin
like
really
just
um
Topic 14
acting
insulin
hours
peak
rapid


count
Gender
views
Gender
Man
No Actor
Female
Medical_info
high
low
Race
asian
indian
white
No Actor
middle eastern
black
latino hispanic

Results


Accuracy score
70
60
50
40
30
20
10
0
Logistic Regression
K-Nearest Neighbors
Decision Tree
Random Forest
XGBoost
Algorithms

Results


Feature importances obtained from Decision Tress coefficients
0.30
0.25
0.20
0.15
0.10
0.05
0.00
likeCount
channelSubscriberCount
commentCount
Age
Race
channelViewCount
Gender
Medical_info

References

1. Liu, X., B. Zhang, A. Susarla, R. Padman, “YouTube for Patient Education: A Deep Learning Approach for Understanding Medical Knowledge from User-Generated Videos,” 2018 KDD Workshop on Machine Learning for Medicine and Healthcare, London, UK, August 2018
2. Halim, Z., Hussain, S., & Ali, R. H. (2022). Identifying content unaware features influencing popularity of videos on youtube: A study based on seven regions. Expert Systems with Applications, 206, 117836.
3. Ladhak, F., Durmus, E., Cardie, C., & McKeown, K. (2020). WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. arXiv preprint arXiv:2010.03093.
4. Liu, C., Wang, D., Liu, C., Jiang, J., Wang, X., Chen, H., ... & Zhang, X. (2020). What is the meaning of health literacy? A systematic review and qualitative synthesis. Family medicine and community health, 8(2).
5. Pitoura, E., Stefanidis, K., & Koutrika, G. (2022). Fairness in rankings and recommendations: an overview. The VLDB Journal, 1-28.
6. Serengil, S. I., & Ozpinar, A. (2021, October). Hyperextended lightface: A facial attribute analysis framework. In 2021 International Conference on Engineering and Emerging Technologies (ICEET) (pp. 1-4). IEEE.
7. Thelwall, M., & Foster, D. (2021). Male or female gender-polarized YouTube videos are less viewed. Journal of the Association for Information Science and Technology, 72(12), 1545-1557.
8. Krishna Pothugunta, Xiao Liu, Anjana Susarla, Rema Padman.On Curating Responsible and Representative Healthcare Video Recommendations for Patient Education and Health Literacy: An Augmented Intelligence Approach (https://doi.org/10.48550/arXiv.2207.07915)

Thank You