Executive summary

Background: Myriad of videos are being uploaded in this era of internet. A lot of those videos are about healthcare topics claiming to provide information about symptoms, cures, post treatments of disease and other health related issues. A lot of videos are genuine, and many others do not provide justice to what they claim to inform you about.

Objective: This project aims at identifying different topics that are available in different YouTube videos that are uploaded in the topics of ‘Brest cancer’ and classify if the videos are genuine of fake about the claims that they make in the video

Methods: Every YouTube video is split into different frames, and checked using OpenCV if a video is animated or if an expert is speaking in the video using image analysis.

Transcripts from videos are extracted if available and if a video doesn’t have transcripts available speech-to-text APIs are used and transcript is extracted. The transcripts are then subjected to text pre-processing activities such as removing stop words, punctuations etc.

Topic modelling using supervised and semi-supervised algorithms is performed and most important topics are identified. Dataset is built using metrics such as upvotes, downvotes, comments, views, DISCERN score from a sample 100 videos.

Topic modelling consisted of preparing a bag of words for 4 different topics of “Symptoms”, ”Prevention”, ”Treatment” and “Remedies” that are picked up from the reference papers. Each transcript is then subjected to supervised modelling where percentage of the transcript covering the 4 topics is shown by the model

Results:

Of the 100 videos being tested 20 of them are detected to be animated and hence removed from being tested for fake videos.

For the rest of the videos which are subjected to topic modelling majority of the videos speak about “symptoms” and “Treatment” topics of regarding breast cancer. 35% of the videos have “Symptoms” as major topic, 25% of the videos have ”Treatment” as major topic, 25% of the videos have ”Prevention” as major topic, 15% of the videos have “Remedies” as major topic

Conclusion:

Of the 100 videos initially selected for the analysis, majority of the uploaded videos contain a speaker/doctor/expert or a patient speaking about his/her experience or advice of expertise related to breast cancer.

Important words of the transcript like ‘breast cancer’, ‘stage’, ‘cell’, ‘treatment’, ‘therapy’, ’category’ , ’risk ’ turn out to be the most significant word when it comes to identifying videos which are not fake.

Majority of the videos contained words belonging to topics of “Treatment” and “Symptoms ” of breast cancer using Topic modelling.

Around 60 of the 80 non-animated videos contained information majorly regarding Treatment or Symptoms leading us to conclude that most of the videos are informative.

The unexpected finding that we came across while doing this project is that, fake videos contained higher number of likes compared to informative videos which goes on to say that fake videos spread faster that informative ones.