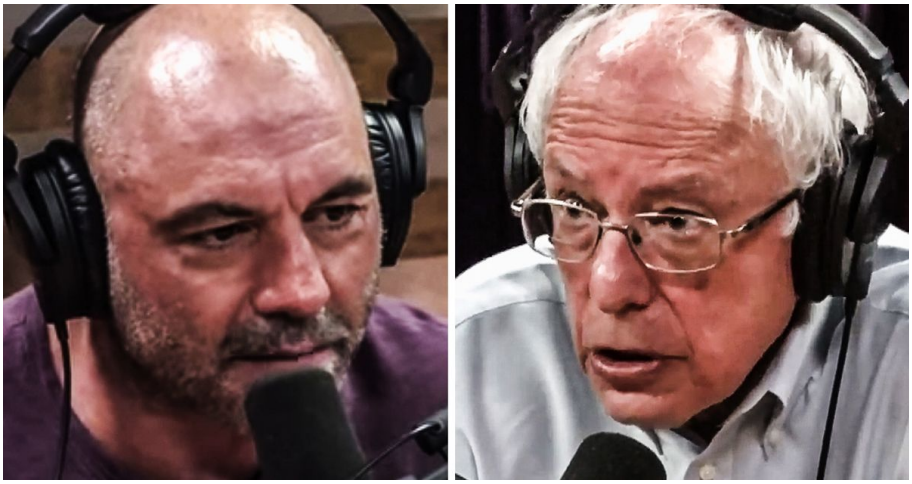
YouTube Comment Sentiment Analysis and Dislikes: Is There a Correlation?

The Joe Rogan Experience is a podcast hosted by MMA commentator and comedian Joe Rogan. The show regularly has guests on who range from athletes and comedians to scientists and politicians. The JRE (Joe Rogan Experience) is among the longest standing and most viewed podcasts in the world. It is available on platforms such as YouTube, Spotify and Vimeo.

We will use the comments from the JRE YouTube videos to determine if there is a correlation between the sentiment of the comments and the like/dislike ratio.



(Joe Rogan Host - left, Bernie Sanders presidential candidate - right)

Our Aim

In our project, we aim to analyse YouTube comments. We are interested in whether there is a relationship between the sentiment of YouTube comments and the like/dislike ratio. We may also explore other relationships with the YouTube comment sentiment, such as views, type of guest (sports, scientist, politician), most liked comment, and longest comment threads. There are many interesting types of data we can analyse.

Our Dataset

The datasets we will be using is data scraped from the YouTube API. The format of the data is in JSON. In the API documentation, we can use a 'GET' request for a variety of data. The relevant ones for us will be getRatings, comments, commentThreads, and videos, and perhaps some more if we find it to be relevant to our analysis. Here's a link to the relevant documentation page, to show the format of the JSON GET responses:
https://developers.google.com/youtube/v3/docs/comments#resource-representation
https://developers.google.com/youtube/v3/docs/videos/getRating#response

https://developers.google.com/youtube/v3/docs/videos#resource
As each of these are separate JSON formats, we will have to join them together. There may be some rows with missing values. We may drop those rows or impute the values manually. Linking these different datasets will be necessary to the analysis we want to perform. It is particularly necessary to join getRatings and comments, as joining these will be the only way for us to use them both together.

Analysis Techniques

The techniques we plan to use to analyze our data set may include, KNN, Kmeans, logistic regression, naive bayes classifier, and SVM.

We expect to use Regular expression techniques to format the comments for processing and for the sentiment analysis we expect to use naive bayes classifier. To predict the likes/dislikes ratio, we expect to use logistic regression, or clustering techniques. We want to use logistic regression on the different categories of comment sentiment (x) to predict the likes/dislikes rating on the video (y). We may use clustering techniques to predict the kind of guest on the episode (comedian or MMA fighter or political figure).

Milestone 1: Preparing and formatting our data for analysis

- Retrieve our data via the YouTube API.
- Join our datasets together.
    - getRatings JSON, comments JSON, commentThread JSON, videos JSON.
- Process our data into a ‘bag of words’ form for analysis.
- Format our rows of data.
    - Remove outliers and rows with NaN values.

We aim to achieve this by the end of week 8, as well as get all of our data from the YouTube API, load them into dataframes, then join them together, and clean them up.

Milestone 2: Apply sentiment analysis on all of our comment data

By week 10 we aim to achieve our sentiment analysis predictive model on our comments. Our model will predict whether a new comment is negative or positive, and possibly other categories such as political, or sports related etc. After this, we will begin to use our sentiment analysis of comments to determine the relationship with likes/dislikes.