Choice of dataset

I plan on using this dataset to train my model. I chose this dataset as it already contains the necessary data fetched from Youtube comments, such as the like count, view count, and replies. Therefore, this dataset would be optimal for training my model. If I want to expand the dataset, I can add more comments using the Youtube API. The disadvantage of this dataset is that there is no label for the sentiments. Therefore, I would most likely need to use an unsupervised model.

Methodology

a. Data Preprocessing: The most helpful information provided by this dataset is the comments derived from real Youtube videos. From these comments, I could use them for training the model. To preprocess the data, I can remove punctuations from the comments, change all comments to lowercase, remove stop words, and stem the words.

b. Machine learning model: I would like to predict youtube videos' overall sentiment/reaction from the comments from the dataset. The videos are categorized into three classes: positive (happiness, excitement, contentment), neutral, and negative (sadness, anger, frustration). If I use the dataset above unsupervised, I will use the K-means clustering to classify the comments in their class. Then, the class with the most data points would be the overall sentiment related to a specific video. The advantage of using K-means is that it is simple and efficient since it can arrive at an optimal solution quickly (fast convergence). The cons of K-means are that the k value is sensitive and needs to be defined appropriately and that we can get clustering outliers which can lead to incorrect predictions. If I were to use a supervised dataset, an alternative model I can use is the Naive Bayes Classifier to estimate and predict the overall feedback of the video. The Naive Bayes Classifier can be helpful as it would analyze each word and calculate the probability of each comment belonging to the three different categories by analyzing the word combinations. The category with the highest probability would represent the overall sentiment. The pros are that Naive Bayes Classifiers are efficient and straightforward to implement. The cons are that when this model encounters an unseen feature, it can estimate the probability to be zero, which can be an issue as comments can be pretty varied.

c. Evaluation Metric: I plan on using the Silhouette Score to evaluate the accuracy, as I would most likely go with unsupervised learning. Using the Silhouette Score, it would calculate a score for each data point. A Silhouette Score of 1 would mean that the clusters are well-separated and optimal, -1 implies that clusters are not assigned correctly, and 0 means that the clusters are indifferent. The model should predict the correct sentiment on the Youtube video based on the comments with 70-80% accuracy.

Application

I plan to integrate my model into a chrome extension or a web application. In the chrome extensions case, when you click on a video, it will read the video's id from the URL, fetch the comments related to the video, then perform the necessary analysis.