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.

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 one of three classes: positive (happiness, excitement, contentment), neutral, and negative (sadness, anger, frustration). I would use logistic regression to classify in which category the video should be. Furthermore, this model usually analyses past data to predict the current input by learning the general relationship between feature and dependent variable, which can be helpful as comments can be varied. However, multinomial logistic regression is sensitive to outliers, which can lead to incorrect predictions. 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 confusion matrix and accuracy loss to evaluate if the model categorizes the videos in the correct sentiment category. Using this evaluation metric, it would compare the actual class of the sample to the predicted class, providing statistics on true positives, false positives, true negatives, and false negatives. 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 and fetch the comments related to the video to perform the necessary actions.