Philip Tam

Project Deliverable 2

1. Problem Statement

Synopses or summaries are meant to indicate the genre of a movie. This project will attempt to do genre classification for movie from synopses or summaries using the title and summary.

2. Data Preprocessing

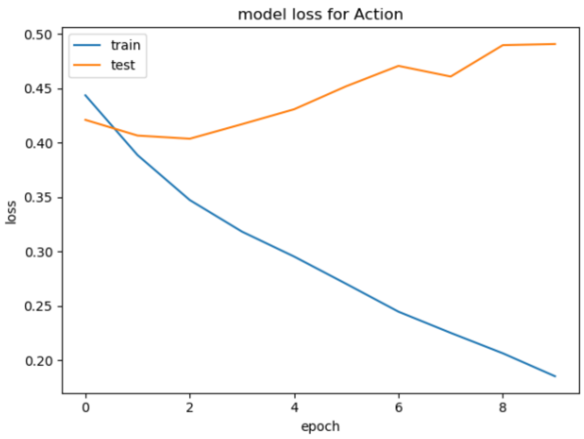
The CMU Movie Summary Corpus Dataset from Bamman, O'Connor, & A. Smith (2013) is the main dataset that has been used. The plot summaries were extracted from the *plot_summaries.txt* and then cleaned to remove all special characters and noise, and lowercased. Each plot summary will be split into different strings with the delimiter period. The titles and genres were extracted from the *movie.metadata.tsv*. There are over 366 genres where 8 were selected because of their count and uniqueness: Action, Comedy, Adventure, Romance, Crime, Horror, Thriller, and Drama. Each of these genres have at least 3000 movies that are associated with it respectively. Since movies can be multi-genre, movies will be able to have multiple different genres. To select the features, the summaries have to be vectorized. For this preliminary stage, a method of taking the 10000 most common words from the summaries will be where the length of each sentence will be 500, and then tokenizing the sequence with Keras' tokenizer.

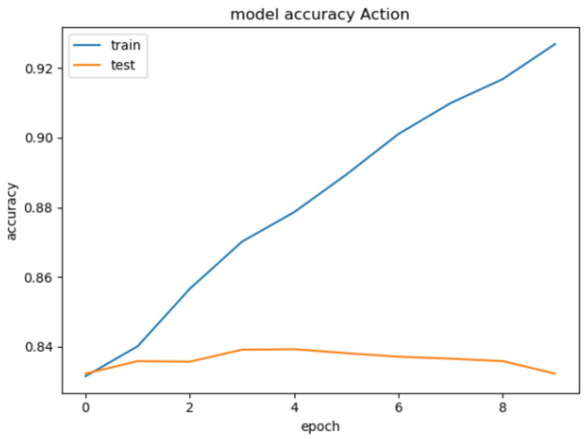
3. Machine learning model

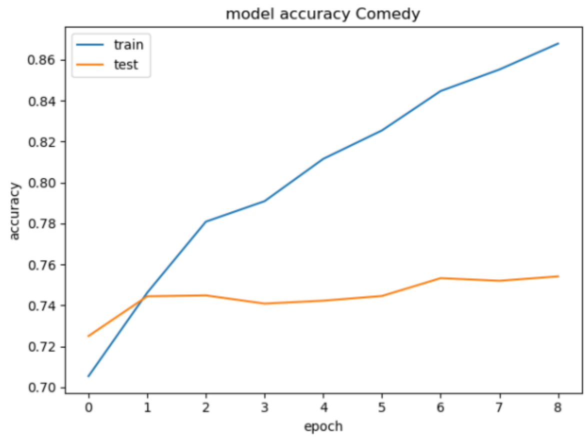
The main framework will be Keras because of its simplicity. The main model of this project will be a Bidirectional LSTM model similar to Ertugrul's (2018). The dataset splits are 80% for the training set and 20% for the test/validation set because more data should provide would better results. For this preliminary stage, other models will be used such as support vector machine,and logistic regression to compare to the bidirectional lstm using a onevsall model because it provides better results than if all of the classes were trained simultaneously. The other models will be using the Tidfvectorizer to find the features. The batch size will be will be 32 because of the large amounts of data. The model ran through 10 epochs because it is standard. The loss function is a binary crossenthropy because the classifications are all done individually. The optimizer is adam because it is model that changes its learning rate dynamically. Moreover, the activation function is sigmoid because it is a classification problem.

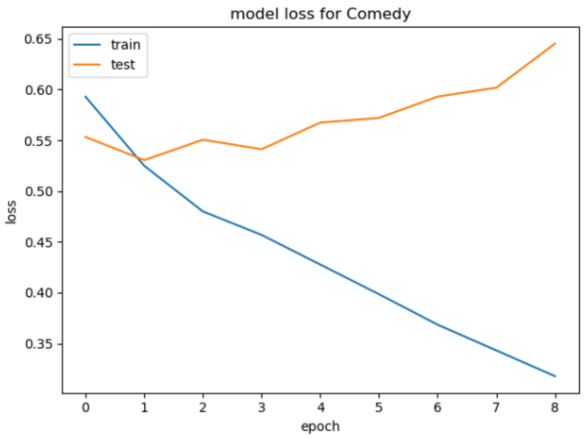
4. Result

The results indicate overfitting because as the improvement in the train accuracy increases as well as the loss, test loss increases and test accuracy stagnates for both Action and Comedy.

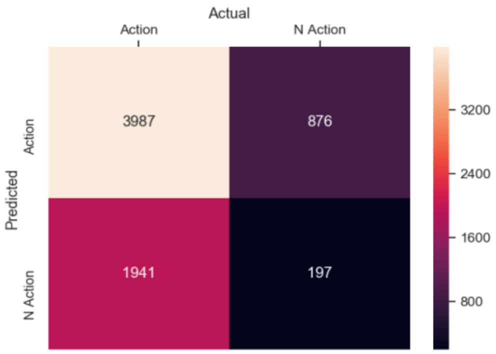


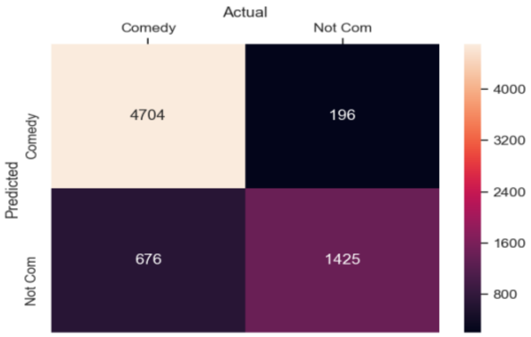




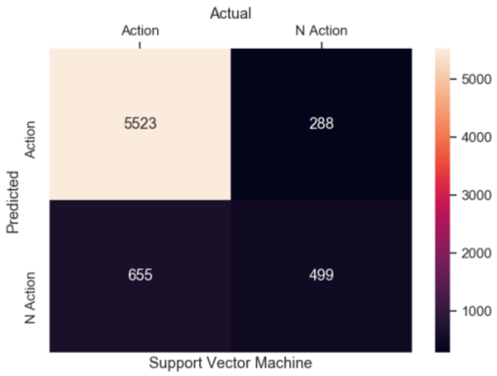


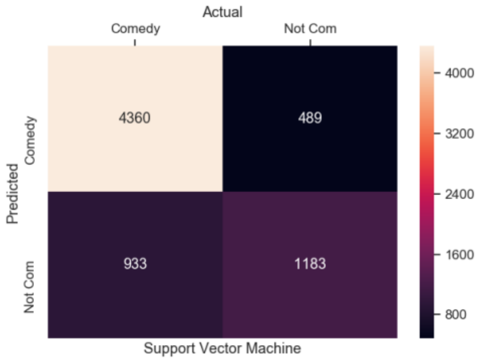
Moreover, the confusion matrix of action of the bidirectional lstm model indicates that the model has cannot finding patterns in the action model, while the comedy model seems to fair quite better. Note: Action should be non-action, and non-action should be action (i.e. reverse the label names.)

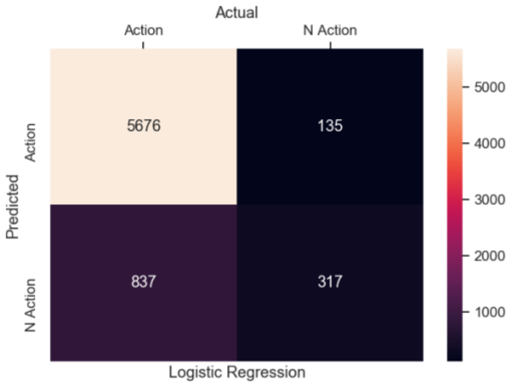


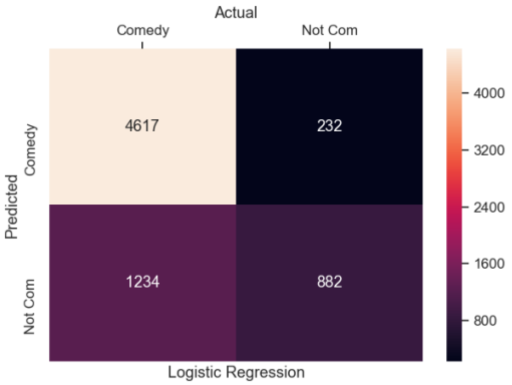


Additionally, the bidirectional lstm model performs much worse than the support vector machine model for action, and for comedy. This is also true for the logistic regression model.









5. Next steps

To further improve the performance of this model, the main solution would be to acquire more data from other sources since the main issue is the lack of data. Other methods to improve the model revolve around improving the data processing such as feature extraction method. Other methods to extract features from the summaries will be used such as word2vec, doc2vec, skip-gram model, or even sklearn's Tidfvectorizer. It is also possible to tweak the parameters of the model and adjust them to improve the model's performance such as the batch size, epoch, etc. Other ways of improving the model can be by cleaning the data properly or limiting the number of genres to predict.