Project Progress Report

Text Information Systems (Fall 2021)

Free Topics | Sentiment Analysis and Topic Modelling on Movie Reviews

Tasks Completed:

- We will be using the IMDB dataset from Stanford, we tried to explore some datasets from different websites but we did not get any other labelled datasets. We found something from cornell but the pre-processing for that dataset will take a lot of time because that is just a scraped dataset with a lot of unnecessary data.

- Then we have divided the dataset into train, test and validation datasets.

- There were some cleaning and preprocessing required in the dataset like removing stop words, adding sentence start and end symbols, converting a review into tokens etc.

- So far, we have created two models, one is CNN and the other is RNN, but training and testing is still left.

- Also, for validation part, we have scraped the the IMDB movie reviews from their website and we have scraped reviews for more than 100 movies.

S.N	Task	Estimated Time (in hours)	Status
1	Explore the datasets	6	Done
2	Combine all the labelled datasets of movie reviews	2	Done
3	Pre-processing on the dataset	2	Done
4	Explore different Machine Learning and Deep Learning algos to create the classifier	15	Partially Done
5	Scrape the real time movie reviews for some movies for the evaluation from IMDB website	5	Done

Challenges:

We have not faced any serious challenges so far. But, we did face some challenge while exploring different movie datasets and deciding which dataset we will be using. It took more than expected time.

Then, there would be one more challenge when we will be deciding which algorithm works best for classifying the sentiments of movies. We also, faced one challenge that out dataset just has reviews and its sentiment but not the movie name. So for topic modelling if we have movie name then it could be a better deal.

Tasks Pending:

- We need to explore different algorithms both machine learning and deep learning algorithms to understand which models works best.
- Then, we need to start doing the topic modelling on the reviews.
- Then, we have to train our model and do the basic testing on testing dataset.
- To Compute error metrics and do testing of code.
- We need to create an API for the classifier.
- We also need to manually review the sentiments of reviews that we have scraped for better testing of model.

S.N	Task	Estimated Time (in hours)	Status
1	Explore different Machine Learning and Deep Learning algos to create the classifier	15	Partially done
2	Topic modelling on the whole dataset	10	Not done
3	We will manually find the sentiment from the web for the above scraped movies for evaluation	5	Not done
4	Compute the error matrices	4	Not done
5	Testing and debugging	6	Not done
6	Create an API	5	Not done

Changes in Proposal:

There were few comments on our proposal and it seems we were unclear that from where we will going to get out dataset and for manual evaluation, from where we will be getting the dataset.

We have made the changes for the above asked questions and you can review the latest proposal on the same repository.

Now, we will be using the stanford's IMDB movie reviews dataset which has 50000 labelled movie reviews. For evaluation we have scraped the movie reviews from IMDB website which has no labels on the reviews, and we have scraped more than 100 movies.

We will be doing manual evaluation on 10-20 movies to get the accuracy of the model on the more real test data.

Team Information:

S.No	Name	NetIDs
1	Ankur Aggarwal	ankura2
2	Saniya Wanhar	swanhar2
4	Jaskirat Singh Pahwa (Captain)	jpahwa2