Project Proposal

Text Information Systems (Fall 2021)

Sentiment Analysis and Topic Modelling on Movie Reviews

Description:

We are creating the movie reviews classifier, where we will predict the sentiment of a movie based on the text data that we will extract from social media websites, for eg Twitter. Also, we will do some analysis on the sentiments of different people for the same movie. Sentiments analysis plays a very important role for understanding what people, critics etc think about the movie, and it gives a clear picture to the producer, director, cast of the film and other people who are planning to watch a movie a some sense of how the movie is and what people think about it. It plays an important role while measuring the success rate of movies and what are the things that the audience likes and what can be changed. Also, to get a more clear picture we also also do some topic modelling for the sentiments that were positive and also the sentiments that were negative.

Task:

The main goal of this research is to use textual information to determine the underlying attitude of a movie review. In this research, we attempt to classify if a person enjoyed a film based on the review they leave for it. This is especially beneficial when a movie's creator wants to assess the film's overall performance based on reviews from reviewers and moviegoers. Also, after doing the analysis on sentiments, we are planning to do the topic modelling on the reviews to get a better essence of what people think about the movie.

Value:

It will help the cast, directors and people who are planning to watch the movie about the sentiment of that movie on social media which can influence the success rate of a particular movie. This model will currently be classifying the sentiments of movie review, later we can add the capability to classify the sentiment on any product and could be extended for many other use cases. This project's output can also be utilised to develop a recommender, which provides movie recommendations to viewers based on their prior reviews. Another use for this project would be to locate a group of people who like the same movies.

Approach:

1. We will extract labelled movies reviews datasets from different sources on the web.
2. We will combine all the datasets and we will do some preprocessing on the reviews.
3. After data preprocessing we will create a sentiment analysis classifier for the same.
4. Also, we will do topic modelling on the same dataset to get better understanding of reviews.
5. Then we will scrape the social media texts about the movies that the model is not trained on for the evaluation part.
6. We will do the evaluation on the scraped reviews and we will calculate the error rate and matrices.
7. Then we will create a simple API for the testing process, so that other people can also interact with the model.

Tools and Datasets:

Tools:

1. Python as a programming language,
2. Twitter api for scraping text reviews,
3. Machine learning libraries such as nltk, scikit learn etc,
4. Fastapi for api creation.

Datasets:

1. We will be using NLTK's imdb movie reviews dataset.
2. Stanford's - Large movies review dataset: https://ai.stanford.edu/~amaas/data/sentiment/
3. We will also explore some other datasets maybe from kaggle.

Expected Outcome:

We are expecting our model to tell us the sentiment of a movie based on its text reviews and it will be a binary classifier, which will classify the review as positive or negative and it will tell us the most common keywords used for both positive and negative reviews.

Evaluation:

We will evaluate our model on the real data that we will scrape from social media. The results of the model will be verified from the web. We will take 10 movies for the evaluation and we will also search the sentiment or success rate of the movie from the internet and we will use that to get our error metrics.

Team Information:

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

Programming Language: Python

WorkLoad:

S.No	Task	Estimated Time (in hours)
1	Explore the datasets	5
2	Combine all the labelled datasets of movie reviews	5
3	Feature Selection and Feature Extraction (if necessary)	5
4	Explore different Machine Learning and Deep Learning algos to create the classifier	15
5	Topic modelling on the whole dataset	10
6	Scrape the real time movie reviews for 10 movies for the evaluation from social media website	15
7	We will manually find the sentiment from the web for the above scraped 10 movies for evaluation	5
8	Compute the error matrices	3
9	Create an API of the model	7

Total Time Estimated: 70 hours