

PROJECT COMPETITION SYNOPSIS

PROJECT GROUP NUMBER – IV.3

FACULTY MENTOR'S NAME – Professor Bijoly Saha

PROJECT STATEMENT NAME – Movie review classification based on review comments

ALUMNI MENTOR'S NAME – Saurav Mondal

GROUP MEMBERS

ROLL

1. Yogesh Soni	13000218017
2. Triloki Nath Jha	13000218023
3. Preety Mehra	13000218068
4. Poulami Saha	13000218071

- **PROBLEM STATEMENT**

The given problem statement is movie review classification based on review comments. Thus, by using machine learning we want to get the reviews of various users about different movies and then categorize these reviews in accordance with the polarity, i.e., if they liked the movie or they hated it. We want to know whether the sentiment of the user for a particular movie is positive, negative and see whether the machine is able to detect that polarity and evaluate the correct output.

- **AIM**

The main aim of this project is to identify the underlying sentiment of a movie review on the basis of its textual information. In this project, we try to classify whether a person liked the movie or not based on the review they give for the movie. We aim to utilize the actual meaning of the statement in the review to predict it's the overall polarity.

- **IMPORTANCE**

Movie reviews are an important way to gauge the performance of a movie. While providing a numerical/stars rating to a movie tells us about the success or failure of a movie quantitatively, a collection of movie reviews is what gives us a deeper qualitative insight on different aspects of the movie. A textual movie review tells us about the strong and weak points of the movie and deeper analysis of a movie review can tell us if the movie in general meets the expectations of the reviewer.

- **TECHNOLOGY USED** – Machine Learning using Python

- **ALGORITHMS USED** – Existing classification algorithms like Naïve Bayes, Support Vector Machine Algorithm, Logistic Regression etc.

- **PROPOSED APPROACH**

1. Labeled polarity movie dataset has been taken in the consideration which consist of 1000 positive and 1000 negative reviews.
2. Each movie review first undergoes through a preprocessing step, where all the vague information is removed. From the cleaned dataset, potential features are extracted. These features are words in the documents and they need to be converted to numerical format.
3. The vectorization techniques are used to convert textual data to numerical format. Using vectorization, a matrix is created where each column represents a feature and each row represents an individual review.
4. This matrix is used as input to classification algorithm and cross validation technique is applied to choose the training and testing set for each fold.
5. Finally, the output is generated which distinguishes each review based on their polarity.

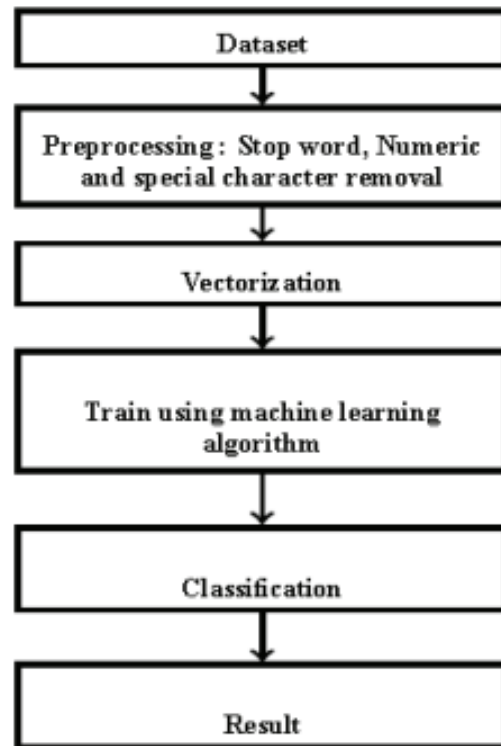


Fig 1. Diagrammatic representation of proposed approach

- **EVALUATION METHODS AND METRICS**

Confusion matrix is generated to tabulate the performance of any algorithms. This matrix shows the relation between correctly and wrongly predicted reviews. In the confusion matrix, TP (True Positive) represents the number of positive movie reviews that are correctly predicted whereas FP (False positive) gives the value for number of positive movie reviews that are predicted as negative by the classifier. Similarly, TN (True Negative) is number of negative reviews correctly predicted and FN (False Negative) is number of negative reviews predicted as positive by the classifier.

From this confusion matrix, different performance evaluation metrics like precision, recall, F-measure and accuracy are calculated.

1. Precision: It gives the exactness of the classifier. It is the ratio of number of correctly predicted positive reviews to the total number of reviews predicted as positive.

$$\text{Precision} = TP / (TP + FP)$$

2. Recall: It measures the completeness of the classifier. It is the ratio of number of correctly predicted positive reviews to the actual number of positive reviews present in the corpus.

$$\text{Recall} = TP / (TP + FN)$$

3. F-measure: It is the harmonic mean of precision and recall. F-measure can have best value as 1 and worst value as 0.

$$\text{F-measure} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

4. Accuracy: It is one of the most common performance evaluation parameters and it is calculated as the ratio of number of correctly predicted reviews to the number of total number of reviews present in the dataset.

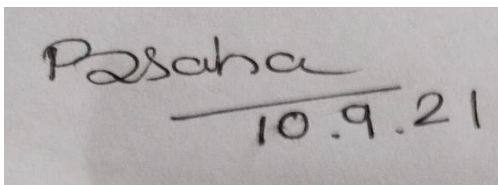
$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

- **CONCLUSION**

The final output will generate whether a review is a like/dislike based on our proposed model. Also, an attempt has been made to classify movie reviews into positive or negative polarity using machine learning algorithms. This model can be used to classify a huge database of movie reviews that will help movie producers to check the status of their movie. Moreover, it will allow them to have an idea on the user's preference of genre.

Hence, this project gave us an opportunity to learn and apply machine learning algorithms to a real-world problem thus gaining knowledge and contributing to the advancement.

FACULTY MENTOR'S SIGNATURE

A photograph of a handwritten signature "P2sahca" in black ink on a light-colored surface. Below the signature, the date "10.9.21" is written, with a horizontal line drawn above it.