Natural Language Processing

Movie Review Analysis

- Project Review Presentation -

This Project Is Done By Sridhar S 2018103068, Jayasurya V 2018103541

CONTENTS

Introduction and Objective

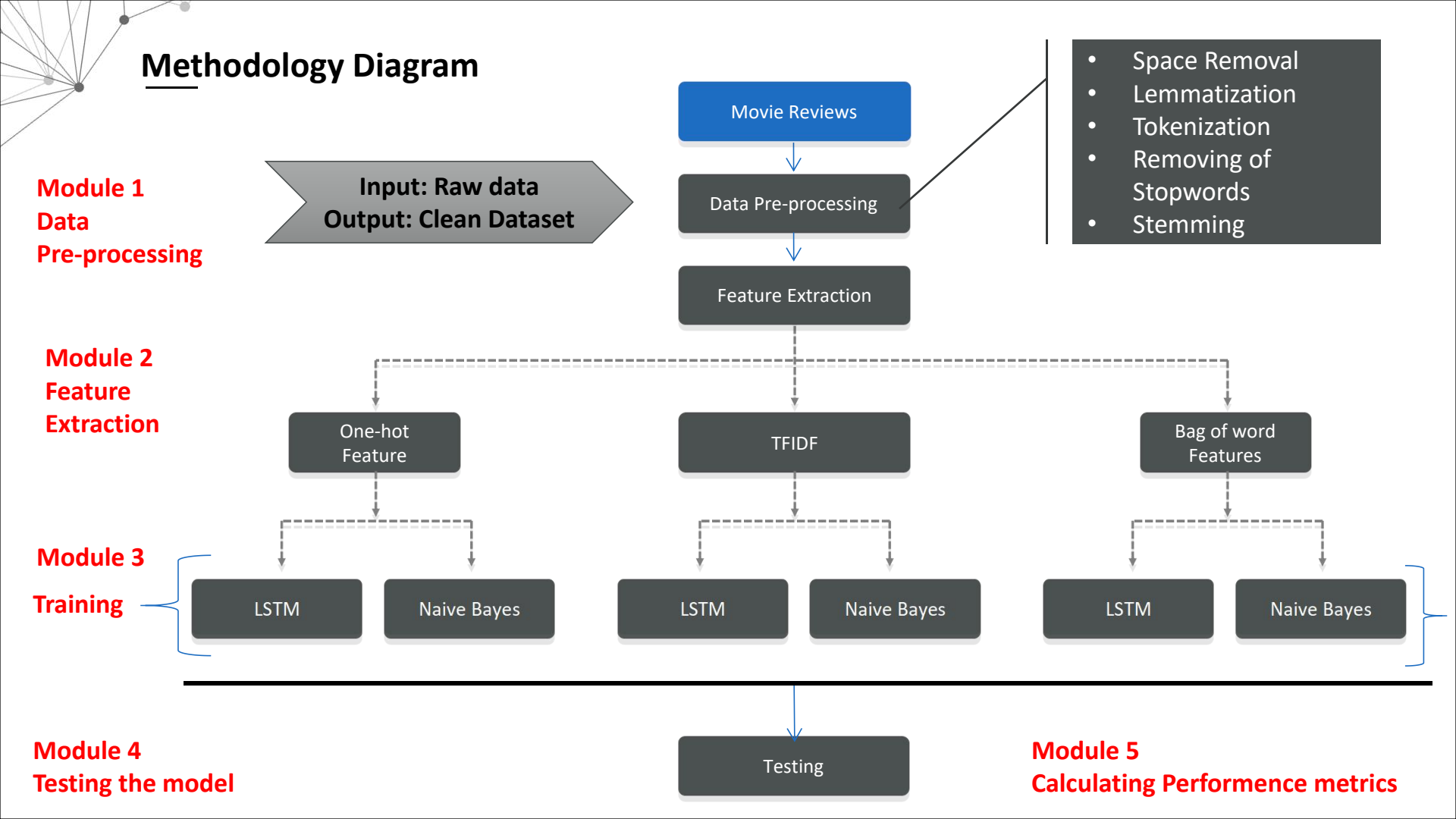
- A movie review is a detailed analysis of a film or a documentary. It involves analysis, research, and reporting the writer's views in a structured way. The writer assumes a position of educating readers whether they have watched the film or not. In fact, many people read movie reviews to decide whether they want to see a film or not.

- The Objective of this Project is to conduct sentiment analysis of movie reviews.The models are to be trained to identify positive reviews and negative reviews.

Dataset Description

- The data was sourced from kaggle.
- IMDB dataset having 50K movie reviews for natural language processing or Text analytics.
- This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets.
- It provide a set of 25,000 highly polar movie reviews for training and 25,000 for testing. So, predict the number of positive and negative reviews using either classification or deep learning algorithms.

Methodology Diagram



Module Description

- **Data Preprocessing**
  - Pre-processing refers to the transformations applied to our data before feeding it to the algorithm.
  - Data Preprocessing is a technique that is used to convert the raw data into a clean data set.

- **Feature Extraction**
  - To classify the text into any category, we need to define some criteria.
  - On the basis of those criteria, our classifier will learn that a particular kind of text falls in a particular category. This kind of criteria is known as `feature`. We can define one or more feature to train our classifier. In this model the features used are **One hot Feature, TFIDF, Bag of words** Feature.

- **Training and Testing Classifier**
  - From the feature set we created above, we now create a separate training set and a separate testing/validation set.
  - The train set is used to train the classifier and the test set is used to test the classifier to check how accurately it classifies the given text.
  - In this model the training algorithm used are **LSTM and Naive Bayes**.

- **Calculating Performance metrics**
  - Performance metrics are known as numbers and data representing Project's abilities, actions, and overall quality.

Performance Metrics

- Confusion Matrix: The Confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model. It is used for Classification problem where the output can be of two or more types of classes.

- Accuracy: Accuracy in classification problems is the number of correct predictions made by the model over all kinds predictions made.

- Loss curve:
  - The Training curve shows the loss or metric calculated on the training subset.
  - The Validation curve shows the loss or metric calculated on the validation subset.

THANK YOU