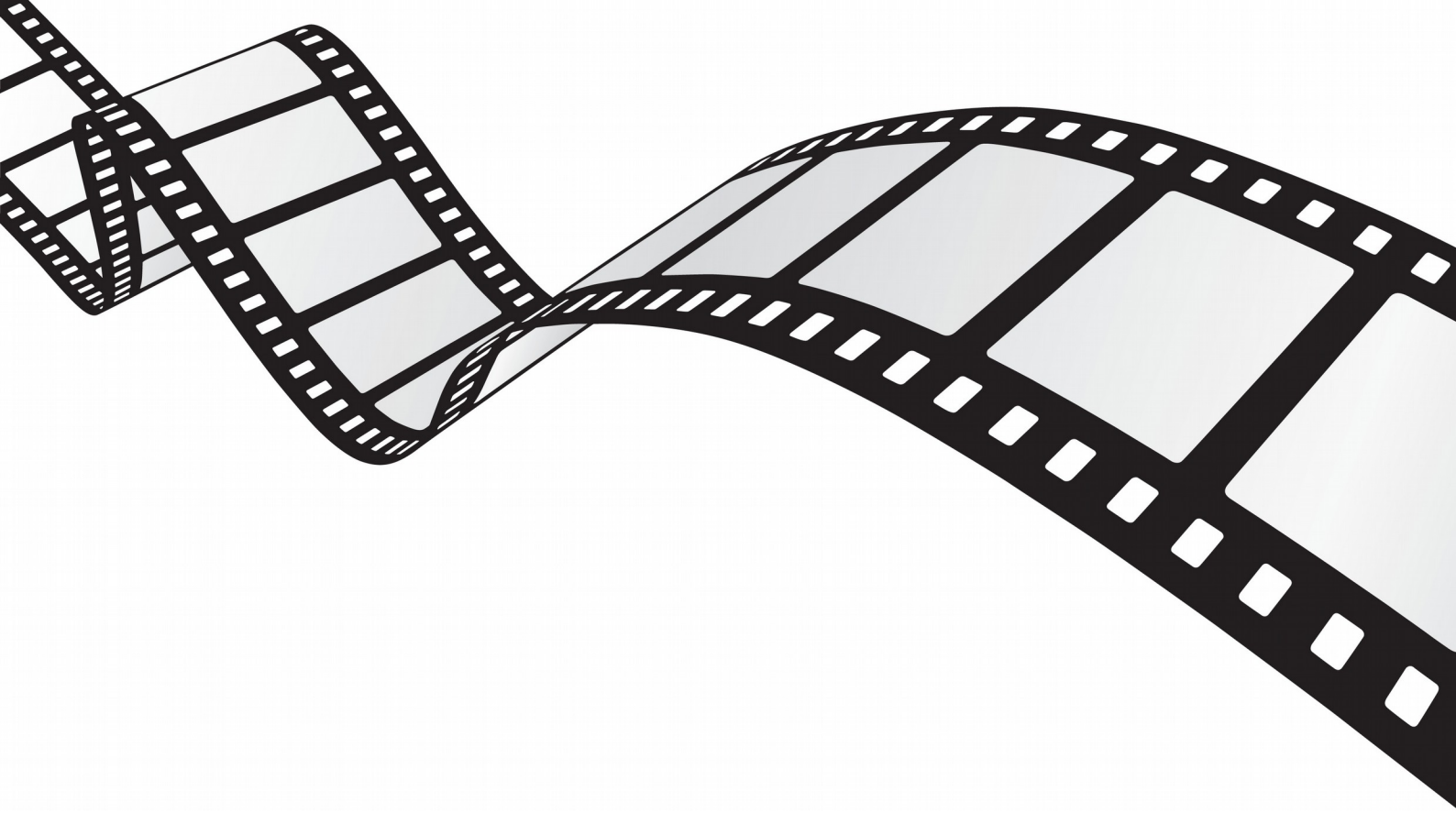

Natural Language Processing of IMDB Movie Reviews

Submitted By : Sumith T S & Sreelekshmi S

Data Overview

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, providing a set of 25,000 highly polar movie reviews for training and 25,000 for testing. Repository located at the following URL:
https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.

Goal

1. Predict the number of positive and negative reviews.

Specifications

1. review(string)
2. sentiment(string)

Methodology

- Pre processing
  - Attribute selection
  - Converting to lower case
  - Removing html tags
  - Removing special special characters
  - Lemmatizing
  - Removing extra white spaces
  - Training and Test data
  - Feature Scaling

- Processing
  - Processing is applying different algorithms to the data to find the best results

Algorithms used

1. Logistic Regression
2. Random Forest Classifier
3. Ridge Classifier
4. Naive Bayes
5. XGBoost
6. Logistic Regression with SMOTE

Results

The data set used for is further splitted into two sets consisting

of two third as training set and one third as testing set. Among the two algorithms applied random forest shown the best results. The efficiency of the two approaches is compared in terms of the accuracy.The accuracy of the prediction model/classifier is defined as the total number of correctly predicted/classified instances.Accuracy is given by using following formula:

Accuracy= (TP+TN/TP+FN+FP+TN)∗100

where TP, TN, FN, FP represents the number of true positives, true negative, false negative and false positive cases.

we can see the prediction and say that Logistic Regression with SMOTE is better perform then logistic regression model.

and the accuracy in Logistic Regression with SMOTE is 90.16