Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Artificial Intelligence Lab

CSE 4108

Sentiment Analysis on Book Review from Bengali Text

Submitted To

Mr. Faisal Muhammad Shah
Associate Professor
CSE, AUST

Mr. Md. Siam Ansary
Lecturer
CSE, AUST

Submitted by

Atanu Kumar Saha	170204003
Md. Mahiat Mahin Apu	170204006
Faisal Ahmmed Tonmoy	170204025

Contents

1 Project Description

Our project titled Book Review Analysis From Bengali text Data will help to detect the polarity of reader's review. For our work, we used various types of Machine learning Models. For each model, we will create a confusion matrix, and then compare the test results.

2 Our Dataset

For our dataset, we collected reviews from various types of online book selling sites, including rokomari, boiferi, and others.It contains 1162 features with more than 700 samples. We considered 80% of data for the training set and 20% for the test set.

3 ML Model

We have used 5 ML classifier models in our project to develop our dataset. A short brief of the models are given below :

- **Decision Tree :**

  Decision Trees(DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Decision Tree can be seen as a piecewise constant approximation. It is a tree structure classifier where a decision node contains features of the dataset and splits data over a decision rule. And Leaf nodes contain the outcome. Leaf nodes also help to decide the class of new data. This model selects features by Information Gain techniques. Model will split every possible condition and take the split with maximum information gain. It needs entropy to calculate information gain.

- **Random Forest :**

  Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. It is based on ensemble learning. It is a collection of multiple random decision trees on various subsets of the given dataset. And takes majority votes to predict the test data of the dataset. In this model the number of subset features = square root of total number of features.

- **KNN (K-Nearest Neighbor) :**

  The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It is one of the simplest ML models. First it chooses k then calculates distance of test data from all training data. Then it considers majority labels of k closest neighbors. And categorized the test data into that label.

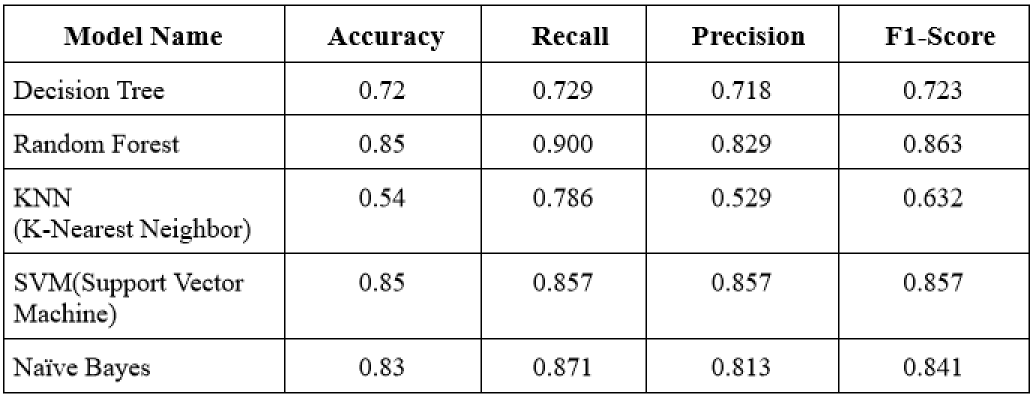
- **SVM (Support Vector Machine) :**

  It is a supervised learning Classifier.It chooses the extreme points that help in creating the hyperplane.The goal of SVM algorithm is to create best hyperplane boundary so that we can put the new data in the correct category in the future.It is effective in high dimensional spaces where the dimensions number is greater than number of samples.It doesn't perform well in case of large datasets.It uses a subset of training points in the decision function so it is also memory efficient.

- **Naïve Bayes :**

  Naive Bayes algorithm is a supervised learning algorithm,which is based on Bayes theorem and used for solving classification problems.It includes high dimensional training dataset.It is the simple and effective algorithm that can make quick predictions.It works on the basis of the probability of an object.It perform well in case of categorical input variables compared to numerical variables.

4 Comparison of the Performance Matrix



Model Name	Accuracy	Recall	Precision	F1-Score
Decision Tree	0.72	0.729	0.718	0.723
Random Forest	0.85	0.900	0.829	0.863
KNN (K-Nearest Neighbor)	0.54	0.786	0.529	0.632
SVM(Support Vector Machine)	0.85	0.857	0.857	0.857
Naïve Bayes	0.83	0.871	0.813	0.841

Figure 1: Performance Matrix

5 Discussion

After fitting our dataset to several models, we can find that KNN models have lower performance metrics than other models. However, the performance metrics of the remaining four models worked well. We may conclude that our dataset works well in Decision Tree, Random Forest, Naive Bayes, and SVM models having accuracy 72%, 85%,85% and 83% respectively. Random Forest(SVM) gave us the maximum accuracy which is 85%.

6 Project Contribution

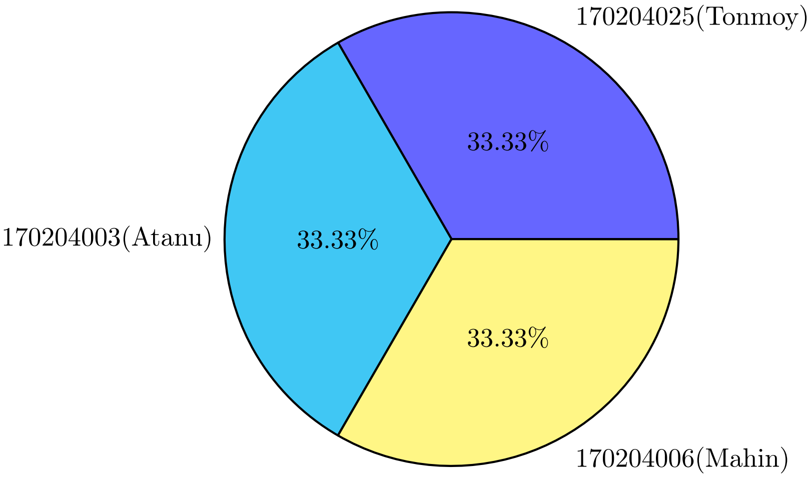


Figure 2: Pie Chart