Pattern Recognition and Machine Learning

Course Project Report

Project Name : Twitter Sentiment analysis

Team Members:

Hari Bhutanadhu-B19EE017

Gattu Hemanth-B19EE030

Guvvala Sujitha-B19EE033

Data Preprocessing and Exploratory Data Analysis :

The task of this project is to build a model which determines the tone (positive or negative) of the texts .The dataset contains target ,id,date,query,username and text columns.

Target column is the class variable .Remaining columns are features .

This is a binary class dataset and each class has an equal number of samples.

We had considered only the text column as the main feature to determine the class ,so dropped the other features from the dataset.

The data we extracted is raw tweets. It contains unnecessary data that is not required for analysis. Now before sentiment analysis, cleaned the dataset by removing double spaces,hyphens ,arrows,URLs,emojis,non english symbols and stop words (Stopwords are those words that do not provide any useful information to decide in which category a text should be classified such as preposition). Converted texts into lower case letters and did spell correction

Counted the most common words frequency,most common negative and positive words frequencies for data visualization

Plotted the word cloud for negative and positive class.Word Clouds are used for visualization of words frequencies in text documents. Word cloud produces an image with frequently appearing words in the text document, where the most frequent words are shown with bigger font,and less frequent words with smaller font.

Classifiers Used :

Logistic Regression:

Logistic Regression is a Machine Learning algorithm which is used for the classification. It is used to predict the categorical dependent variable using a given set of independent variables.

The cost function for hypothesis used is ‘Sigmoid function’ or also known as the ‘logistic function '.In order to map predicted values to probabilities, we use the Sigmoid function.This function maps any real value into another value between 0 and 1.

Sigmoid function is $\sigma(z) = \frac{1}{1-e^{-z}}$

The classifier decides the class ,with a threshold value above which we classify values into Class 1 and the threshold value below which we classify values into class 2.

The cost function(represents optimization objective i.e. we create a cost function and minimize it so that we can develop an accurate model with minimum error)is defined as

$$J(\theta) = -\frac{1}{m}\Sigma[y^{(i)}\log(h\theta(x(i))) + (1 - y^{(i)})\log(1\text{-}h\theta(x(i)))]$$

Multi-layer Perceptron(MLP) :

MLPClassifier stands for Multi-layer Perceptron classifier which in the name itself connects to a Neural Network. It is a class of feedforward artificial neural network (ANN).It utilizes a supervised learning technique called backpropagation for training. The major use cases of MLP are pattern classification, recognition, prediction and approximation.It consists of three types of layers-the input layer, output layer and hidden layer

The input layer receives the input signal to be processed .The output layer performs prediction and classification. An arbitrary number of hidden layers that are placed in between the input and output layer are the true computational engine of the MLP. The data flows in the forward direction from input to output layer. The neurons in the MLP are trained with the back propagation learning algorithm. During forward propagation ,for transferring the activation to the output we use transfer functions such as sigmoid,tanh functions.Backpropagation is a mechanism used to update the weights using gradient descent. It calculates the gradient of the error function with respect to the neural network's weights. The calculation proceeds backwards through the network.

LinearSVC :

The Linear Support Vector Classifier (SVC) method applies a linear kernel function to perform classification and it performs well with a large number of samples. If we compare it with the SVC model, the Linear SVC has additional parameters such as penalty

normalization which applies 'L1' or 'L2' and loss function.

Model Building :

Splitted the data into training and testing datasets with test_size =0.2.Built three models using the above mentioned classifiers using training dataset.

Tested the built models using testing dataset and computed the accuracy ,f1 score,recall and precision of each model.

	Accuracy	F1 score	Mean(cv_scores)
Logistic Regression model	0.78	0.78	0.78
LinearSVC model	0.77	0.78	0.77
MLP model	0.76	0.77	0.54

Compared these model performances using box plot .From the box plot we can tell that logistic regression model has highest mean accuracy , so it can be a better model in terms of accuracy while compared to other models.

Feature Selection :

Performed feature selection using SelectKBest method .SelectKBest retains the first k features of x with the highest scores.chi2 is passed as a score function,SelectKBest will compute the chi2 statistic between each feature of x and y (class labels). A small value of score function will mean the feature is independent of y.

Hence ,Performed SelectKBest and got the k best selected features. Using these features built a logistic regression model .

Tested the built model using testing dataset and computed testing accuracy,f1 score and mean accuracy.

	Accuracy	F1 Score	Mean(cv_scores)
Logistic Regression using Original features	0.78	0.78	0.78
Logistic Regression using SelectKBest Features	0.79	0.80	0.79

Compared these two model performances using box plot .From the box plot we can tell that the logistic regression model using SelectKBest features has high mean accuracy , so it can be a better model in terms of accuracy while compared to the model using original features.

Challenges Faced :

Selecting the best classifier was one of the challenges for us ,as we have tried several classifiers to get better accuracy and performance .

To increase the accuracy was also a challenge for us ,for this we have trained the models by changing parameters of the classifiers used .

Choosing the method for dimensionality reduction was also a great challenge for us ,as we have tried different methods to increase the efficiency of the models .

Conclusion :

By comparing every model the logistic regression model using SelectKBest features has high mean accuracy , so it can be a better model in terms of accuracy while compared to the other models.And we have learnt how to improve accuracy by exploring various methods which resulted in gaining lots of knowledge in machine learning.

Contribution of Team Members :

Hemanth Gattu : Data Preprocessing,Data Cleaning ,built a model using Logistic Regression and tried to improve accuracy by changing hyperparameters and standardizing the data.

Hari Bhutanadhu : Data Visualization,built a model using MLP and SVM and tried to improve accuracy by changing hyperparameters ,normalizing the data and other methods.

Guvvala Sujitha : Done Dimensionality reduction by trying different methods such as PCA,LDA and feature selection methods and tried to improve accuracy by standardizing the data and changing different parameters in feature selection methods.

Report was written in the presence of each team member.

Reference Links :

https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

https://www.bogotobogo.com/python/scikit-learn/scikit_machine_learning_Support_Vector_Machines_SVM.php

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html