CSL2050: PATTERN RECOGNITION AND MACHINE LEARNING

TWITTER SENTIMNENT ANALYSIS

- SAUMITRA AGRAWAL
- B22AI054
- **KRISHNA PATIL**
- **B22CS078**
- GOURI PATIDAR
- B22AI020
- **MUKUND GUPTA**
- **B22CS086**
- AAROHI DHARMADHIKARI
- B22AI001
- **HARSHIT GOYAL**
- **B22CS024**

Problem Statement

To perform sentiment analysis on a dataset comprising tweets by applying classical machine learning algorithms (classifiers), and classifying the tweets as conveying positive, negative, or neutral emotions.

Dataset Description

Data Sample:

	textID	text	sentiment	Time of Tweet	Age of User	Country	Population -2020	Land Area (Kmï¿½)	Density (P/Kmï¿½)
0	cb774db0d1	I`d have responded, if I were going	neutral	morning	0-20	Afghanistan	38928346	652860.0	60
1	549e992a42	Sooo SAD I will miss you here in San Diego!!!	negative	noon	21-30	Albania	2877797	27400.0	105

Dataset info:

```
RangeIndex: 31015 entries, 0 to 31014
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   textID             31015 non-null  object
 1   text               31014 non-null  object
 2   sentiment          31015 non-null  object
 3   Time of Tweet      31015 non-null  object
 4   Age of User        31015 non-null  object
 5   Country            31015 non-null  object
 6   Population -2020   31015 non-null  int64
 7   Land Area (Kmï¿½)   31015 non-null  float64
 8   Density (P/Kmï¿½)   31015 non-null  int64
```

Sentiment Analysis Dataset: Features

Continunous features: 'text ID, text, Population, Land Area, Density'

Categorical features with value counts:

```
neutral     12548
positive     9685
negative     8782
Name: sentiment, dtype: int64
```

```
morning    10339
noon       10338
night      10338
Name: Time of Tweet, dtype: int64
```

```
0-20      5171
21-30     5170
31-45     5170
46-60     5168
60-70     5168
70-100    5168
Name: Age of User, dtype: int64
```

```
Afghanistan         169
Ecuador             169
Chile               169
China               169
Colombia            169
                   ...
Singapore           144
...
Slovenia            144
Solomon Islands     144
Zimbabwe            144
Name: Country, Length: 195, dtype: int64
```

Data Preprocessing and Feature Extraction

1. **Exploratory Data Analysis and Preprocessing:**
   - Data Inspection: **No missing** values detected.
   - Outlier Analysis: **Ignored features with outliers** irrelevant to sentiment analysis like '**Population', 'Land Area', and 'Density'.**
   - Categorical Encoding: The 'sentiment' variable encoded into three classes (**negative, neutral, positive**).
   - Text Processing:
     - Removal of special characters and punctuation using regular expressions.
     - Conversion of text to lowercase.
     - Stemming and stopword removal with NLTK.
2. Data Vectorization**:**
   - Utilization of **BERT (Bidirectional Encoder Representations from Transformers)** for dense 768-dimensional word embeddings.
   - Applied PCA and LDA

Implementation of Different Classifiers

- Decision Tree
- Random Forest
- Support Vector Machine
- Naive Bayes Classifier
- Perceptron
- Logisitic Regression

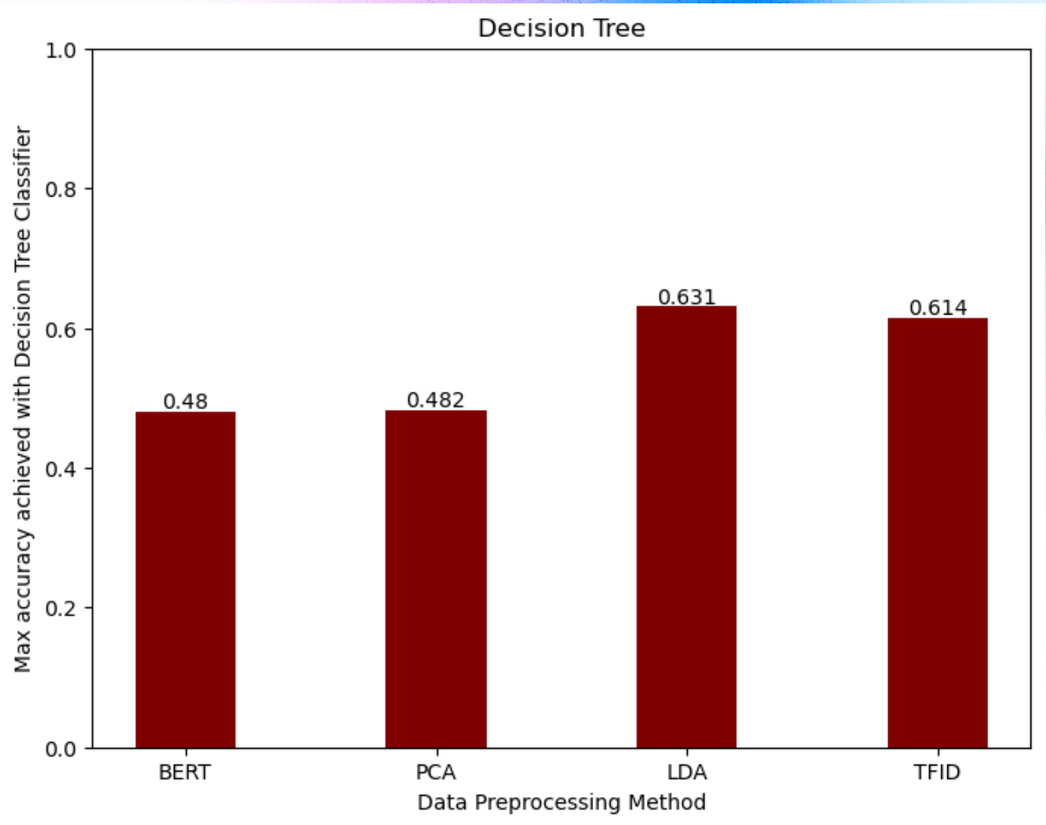
Decision Tree Classifier

OVERVIEW

Implementation: The Decision Tree classifier was implemented using the scikit-learn library's DecisionTreeClassifier model.

Training and Evaluation: The classifier was trained and evaluated across the 4 datasets. the parameter of max_depth was varied and the maximum accuracies obtained for each dataset were recorded and compared.

RESULTS OBTAINED

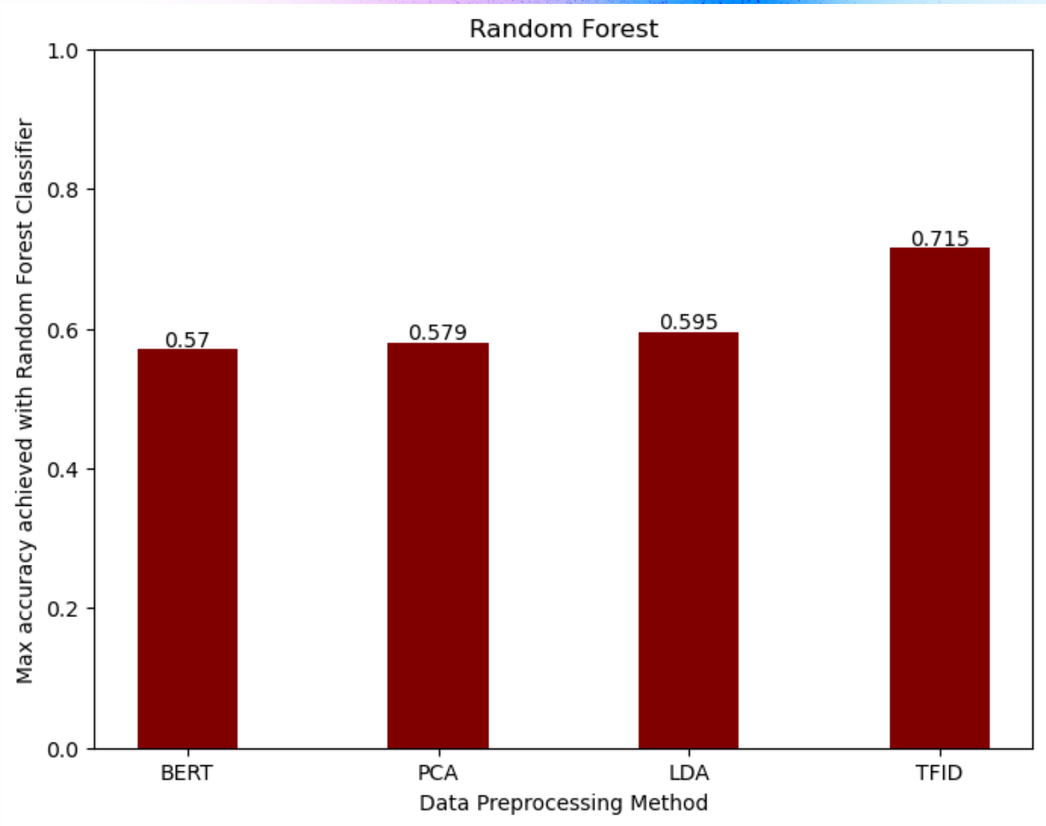


Random Forest

OVERVIEW

Training and Evaluation: The classifier was trained and evaluated across the 4 datasets. The number of trees was varied and accuracies were compared for each value to determin the optimum value for each dataset using gridsearch and then accuracies on each dataset were reported and compared.

RESULTS OBTAINED

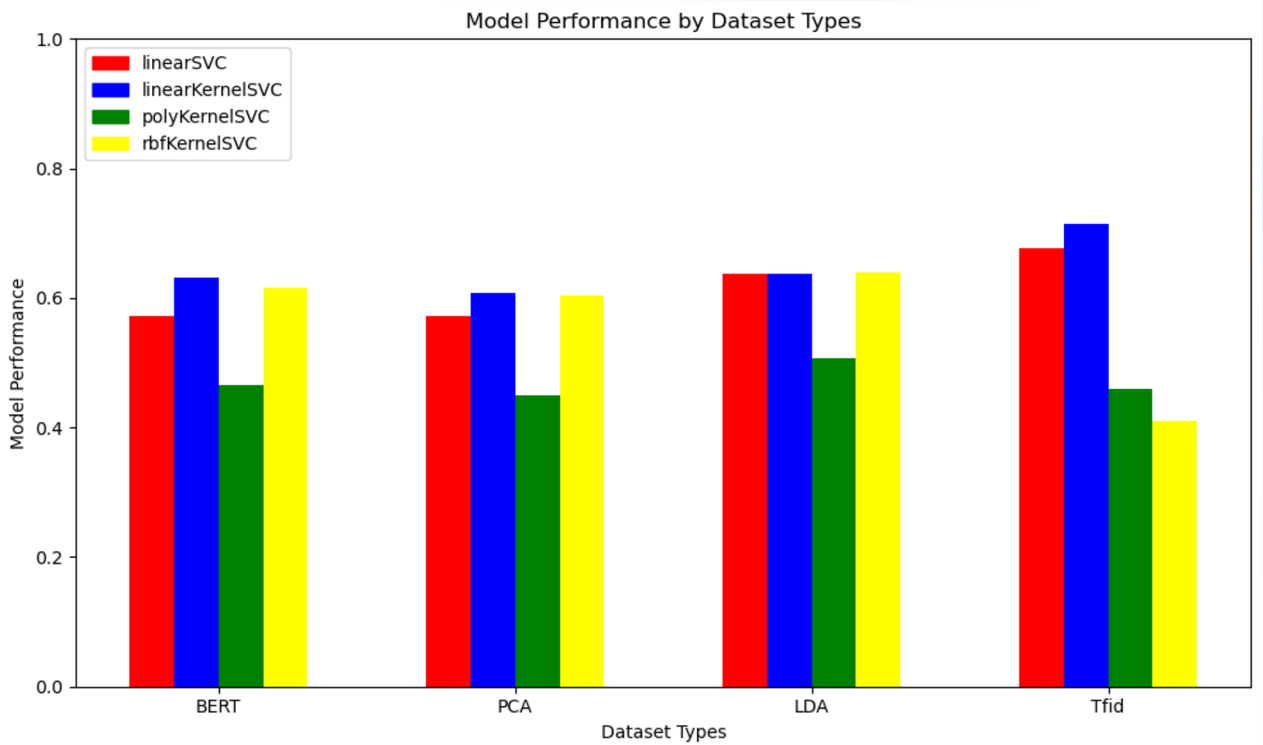


Support Vector Machine

OVERVIEW

Training and Evaluation: To determine the best kernel, 4 kernel settings were tried, for each of the 4 datasets, thereby constituting 16 combinations. Afterwards, the accurcaies and other quality parameters were obtained.

RESULTS OBTAINED

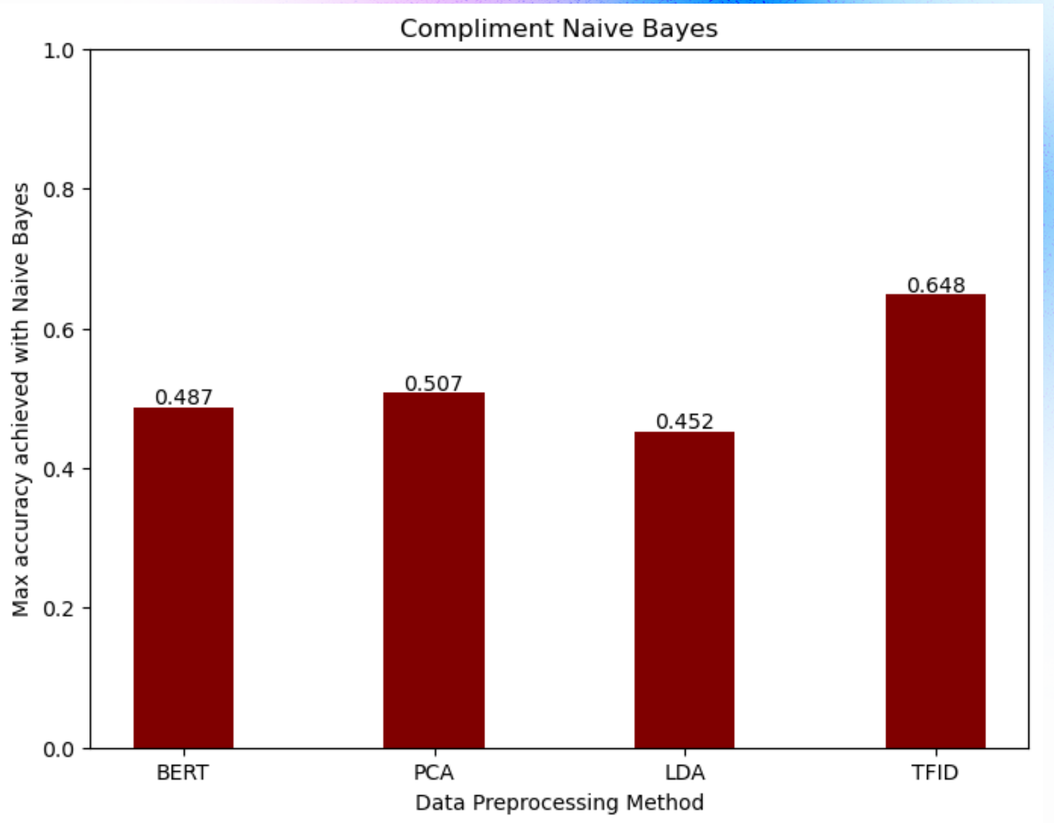


Naive Bayes Classifier

OVERVIEW

Training and Evaluation: The naive-bayes classifier was also implemented using scikit-learn library. Before that, however, the negative values of inputs were corrected. The best hyperparameter value for alpha was determined using gridSearch.

RESULTS OBTAINED

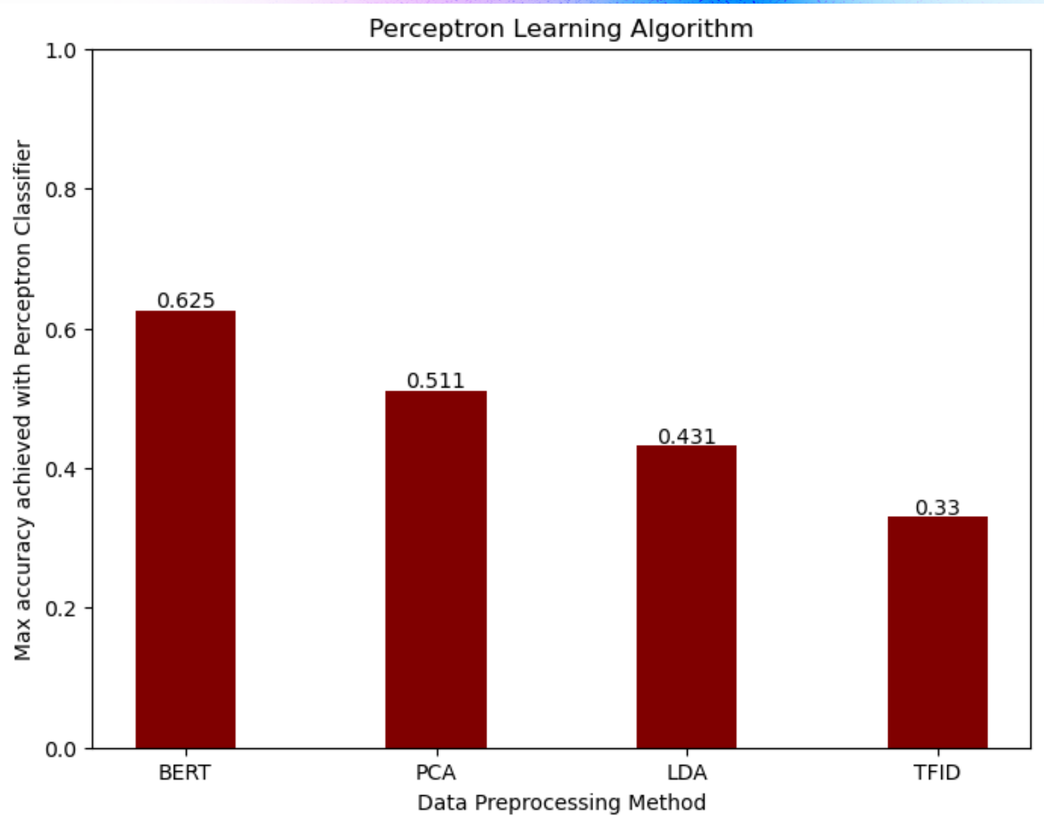


Perceptron

OVERVIEW

Training and Evaluation: We first obtained the optimum values of learning rate and tolerance using gridsearch and then compared the highest accuracies obtained from each dataset.

RESULTS OBTAINED

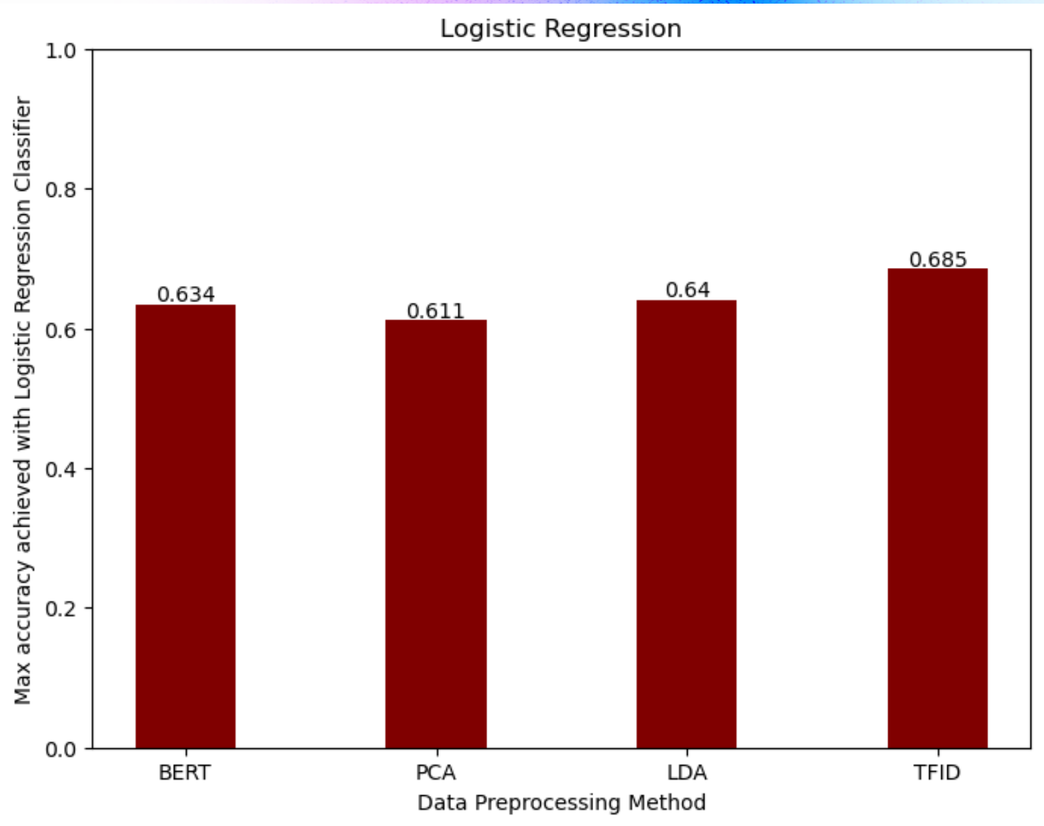


Logistic Regression

OVERVIEW

Training and Evaluation: The optimum value of regulariazation parameter was determined using gridsearch for each of the 4 datasets and the corresponding accuracies were recorded and compared afterwards.

RESULTS OBTAINED



Best performance obtained

Classifier	Dataset that yielded maximum accuracy	Maximum Accuracy
Decision Tree	LDA	0.631
Random Forest	TFID	0.715
SVM	TFID	0.7134
Naive-Bayes	TFID	0.648
Perceptron	BERT	0.625
Logistic Regression	TFID	0.6845

Techniques tried to improve accuracy

01 Ensemble Learning

An ensemble of the top performing models was taken and conflict resolution was done by taking mode of predictions of the 6 models on Tfid vectorized text.

Slightly improved the accuracy and made it stable.

02 Boosting

The incorrectly classified samples were given higher weights then correctly classified samples to further improve the accuracy.

THANK YOU

GitHub Repository

Project Page