Malignant comment classifier project

Submitted by:

Abhilasha mewada

INTRODUCTION

- ## Business Problem Framing

  The proliferation of social media enables people to express their opinions widely online. However, at the same time, this has resulted in the emergence of conflict and hate, making online environments uninviting for users. Although researchers have found that hate is a problem across multiple platforms, there is a lack of models for online hate detection.

  Online hate, described as abusive language, aggression, cyberbullying, hatefulness and many others has been identified as a major threat on online social media platforms. Social media platforms are the most prominent grounds for such toxic behavior.

- ## Conceptual Background of the Domain Problem

  There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influences are facing backlashes from people and have to come across hateful and offensive comments. This can take a toll on anyone and affect them mentally leading to depression, mental illness, self-hatred and suicidal thoughts.

  Internet comments are bastions of hatred and vitriol. While online anonymity has provided a new outlet for aggression and hate speech, machine learning can be used to fight it. The problem we sought to solve was the tagging of internet comments that are aggressive towards other users. This means that insults to third parties such as celebrities will be tagged as un offensive, but “u are an idiot” is clearly offensive.

  Our goal is to build a prototype of online hate and abuse comment classifier which can used to classify hate and offensive comments so that it can be controlled and restricted from spreading hatred and cyberbullying..

- ## Data set description.

The data set contains the training set, which has approximately 1,59,000 samples and the test set which contains nearly 1,53,000samples. All the data samples contain 8 fields which includes‘Id’, ‘Comments’, ‘Malignant’, ‘Highly malignant’, ‘Rude’, ‘Threat’, ‘Abuse’ and ‘Loathe’.

The label can be either 0 or 1, where 0denotes a NO while 1 denotes a YES. There are various comments which have multiple labels. The first attribute is a unique ID associated

with each comment.

The data set includes:

- Malignant: It is the Label column, which includes values 0 and 1, denoting if the comment is malignant or not.
- Highly Malignant: It denotes comments that are highly malignant and hurtful.
- Rude: It denotes comments that are very rude and offensive.
- Threat: It contains indication of the comments that are giving any threat to someone.
- Abuse: It is for comments that are abusive in nature.
- Loathe: It describes the comments which are hateful and loathing in nature.
- ID:It includes unique Ids associated with each comment text given.
- Comment text: This column contains the comments extracted from various social media platforms.

Analytical Problem Framing

- ## Mathematical/ Analytical Modeling of the Problem

  First we have check the null values than describe the datasets

  Than correlation of matrics will be done with the help of heatmap

- ## Data Sources and their formats

  We have used the file train.csv available in the zip folder to train our model as the dataset is available for training.

- ## Data Preprocessing Done

  We have converted all strints into vectors.
  Splitting the strings,
  Combine all threats and negative columns in one columns as a target variable.

- ## Hardware and Software Requirements and Tools Used

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model_selection import train_test_split,GridSearchCV,cross_val_score

from sklearn.preprocessing import StandardScaler,MinMaxScaler,power_transform

from sklearn.linear_model import LinearRegression,LogisticRegression
```

```
from sklearn.metrics import r2_score,
mean_squared_error,accuracy_score,roc_auc_score,roc_curve,confusion_matrix,classification_repo
rt

from sklearn.neighbors import KNeighborsClassifier

from sklearn.ensemble import AdaBoostRegressor,GradientBoostingClassifier

import warnings

warnings.filterwarnings('ignore')

from sklearn import preprocessing

import scipy.stats as stats

from scipy.stats import zscore

from sklearn.svm import SVC

from sklearn.tree import DecisionTreeClassifier

from sklearn.naive_bayes import MultinomialNB

import pickle

from nltk.stem import WordNetLemmatizer

import nltk

import string

from sklearn.linear_model import RidgeClassifier

from sklearn.tree import DecisionTreeClassifier

from sklearn.svm import SVC

from sklearn.ensemble import AdaBoostClassifier

from sklearn.ensemble import GradientBoostingClassifier

from sklearn.metrics import classification_report

from sklearn.feature_extraction.text import TfidfVectorizer
```

Matrices:

RidgeClassifier()

DecisionTreeClassifier()

SVC()

KNeighborsClassifier()

AdaBoostClassifier()

GradientBoostingClassifier()

CONCLUSION

While reading the dataset We find that peoples are criticise the most and all bad comments are than a good one. For our matrics the decision tree classifier give a best accuracy score on training dataset.