***AIM*** : Classifies the user comments in to toxic or non-toxic.

***Dataset*** : jigsaw wikipedia dataset form Kaggle website (Toxic comment - classification challenge).

***Description***:

There are two stages for classifing the comments . This was taken due to the decrease the complexity in the algorithm.

They are :

Stage 1: Rule Based Model

Stage 2: Machine model

***Stage 1: Rule Based Model***

In this Model there is a predefined file which contains list of toxic words i.e frequently used volgure words . So here if the word or words which are in this predefined list are contain in the user comment. Then this model will return “True” indicates comment is a toxic. Otherwise it returns “False” as non-toxic comment.

*Steps for Rule based model:*

---------------------------------

1) Preprocess the data (remove punctuations, remove the numbers , urls etc.....).

Modules : nltk

2) Spell correction

Module : autocorrect

3)Tokinization i.e splitting the sentence in to the list of word.

Module : nltk

4)Each word will check with toxic list .

5) Returns the Results (either “True” or “False”)

***Stage 2: Machine learning Model***

Here we train the model with the given dataset which is jigsaw dataset. The model used for training and testing is Logistic Regression, Naive Bayes MultinomialNB ,

Ridge Classifier, XG boosting. Acutally these four model are comebined for predicting the actual user comments because Precisions of all models are high so combing the models will give better results.

*Steps for Machine Learning Model:*

1) Loading the dataset and Explorate data for analysis . Generally it is a skewed data i.e data

contains very less number of toxic comments compared to the non-toxic commet has ratio

1:10.

Modules : Matplotlib,seaborn

2)Preprocesses the data - remove the punctuations,remove special characters, remove html tags, other url , remove stop\_words, Stemming, Lemmitization.

Modules : re,nltk

3)Convert the preprocessed text in to vector so it is easy to apply mathmetical operations on data. Here we consider the frequently occured words as the features

Modules : tf-idf vectorizer, count vectorizer.

4) Now Pass the word vectorizer in to Machine learning models and train them

Modules :Logistic Regression, Naive bayes MultinomialNB, Xgboost,Ridge Classifier

5) Validation with the dev set .

6) Testing with external user comments.

***Cons :***

***-> This model doesn’t work for the comments the toxic words which includes special characters.***

***Eg: “what the f\_\_\_k”*** *this shows as non-toxic.*

***->It poorly works for the comments which was the context based toxic comments.***

*Generally it will poorly works the comments which are insult or threat comments if it was context based comments.*