

**MALIGNANT COMMENTS CLASSIFICATION**

Submitted by:

Yedhu Chandran

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

**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 behaviour.
* 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 unoffensive, 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.
* Conceptual Background of the Domain Problem

ML can be used to detect the Malignant comments posted by the

Users and take appropriate action

* Review of Literature

The research included comments from threat , loathe, highly malignant , rude , malignant , abuse .

It should the highest comments are from malignant group and rude comes the second

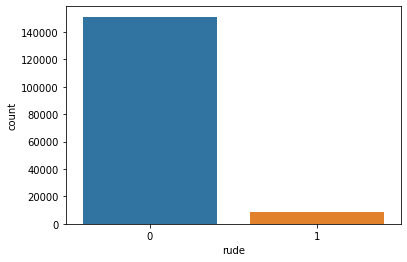
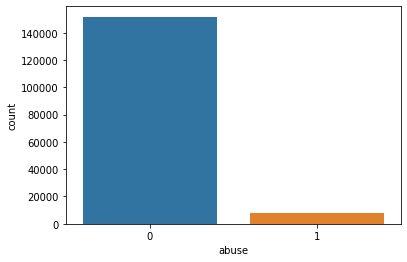
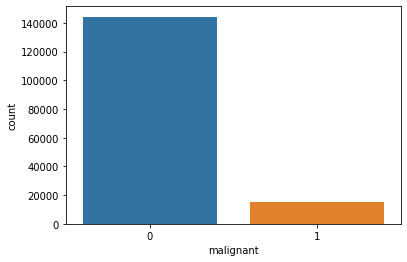
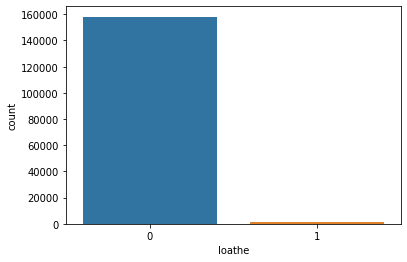
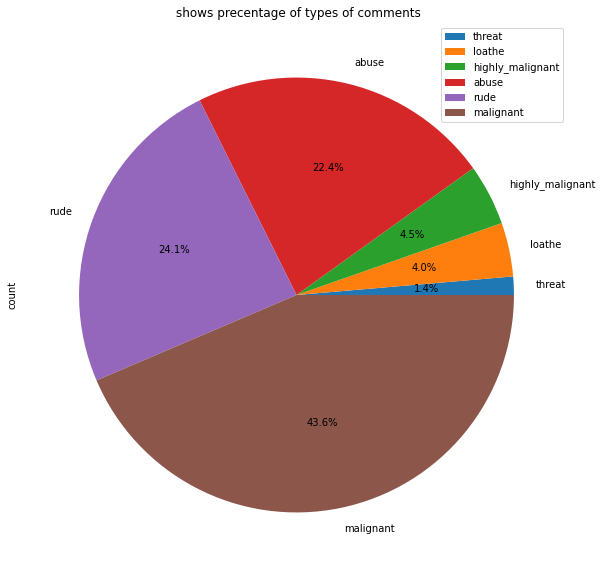
* Motivation for the Problem Undertaken

goal was 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.

**Analytical Problem Framing**

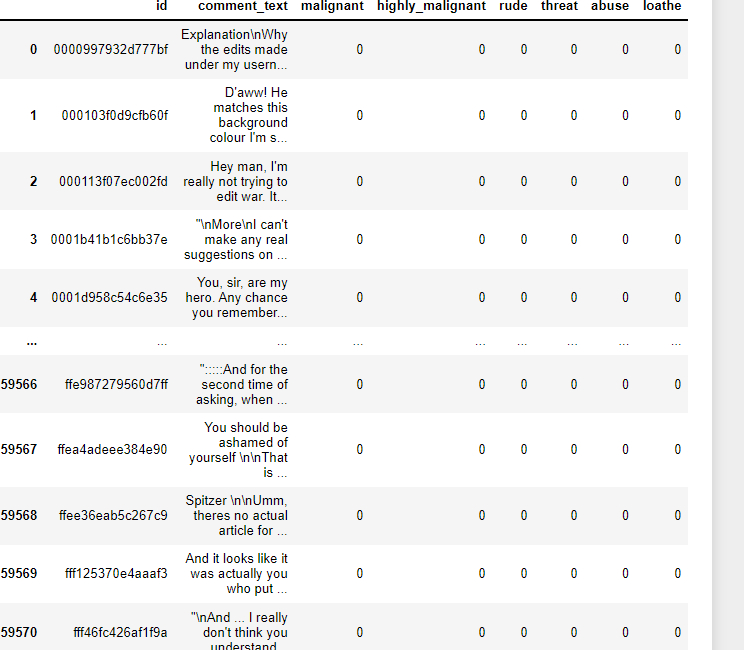
* Mathematical/ Analytical Modeling of the Problem

As we can see from the below chart , highest comments are from malignant and then rude , abuse

* **   **
* ****
* Data Sources and their formats

The data has 159571 rows × 8 columns

* The columns are:
* 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.



In [ ]:

**Data had no null values**

**Correlations were not high**

**Skewness were not there**

**No outliers**

* Data Preprocessing Done

The data was cleaned by replacing the '^.+@[^\.].\*\.[a-z]{2,}$', with 'emailaddress'

'^http\://[a-zA-Z0-9\-\.]+\.[a-zA-Z]{2,3}(/\S\*)?$', with webaddress'

'£|\$', with 'dollers'

'^\(?[\d]{3}\)?[\s-]?[\d]{3}[\s-]?[\d]{4}$', with "phonenumber"

Converted all comments to lower case

Processed using WordNetLemmatizer

And

TfidfVectorizer

* Data Inputs- Logic- Output Relationships

**The input data was the comment\_text , which contain all the comments needed for the model**

**Output was the malignant comment column as w are predicting this .**

**0 is not malignant**

**And**

**1 is malignat**

**The outpt and the input has no correlation**

* State the set of assumptions (if any) related to the problem under consideration

Here, you can describe any presumptions taken by you.

* Hardware and Software Requirements and Tools Used

Tools used are Python notebook ,

Library used are numpy, pandas, seaborn, sklearn,

Wordcloud , WordNetLemmatizer , TfidfVectorizer ,

Logistic regression ,

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

* Testing of Identified Approaches (Algorithms)

Algoriths used are logistic regression

Multinomial NB

Cross validation

Grid search cv

* Run and Evaluate selected models

Logistic resression was the most accuracy model with accuracy for testing 0.9545041516528279

Precision , recall , f3 , support as follows

precision recall f1-score support

0 0.96 1.00 0.98 28856

1 0.93 0.57 0.71 3059

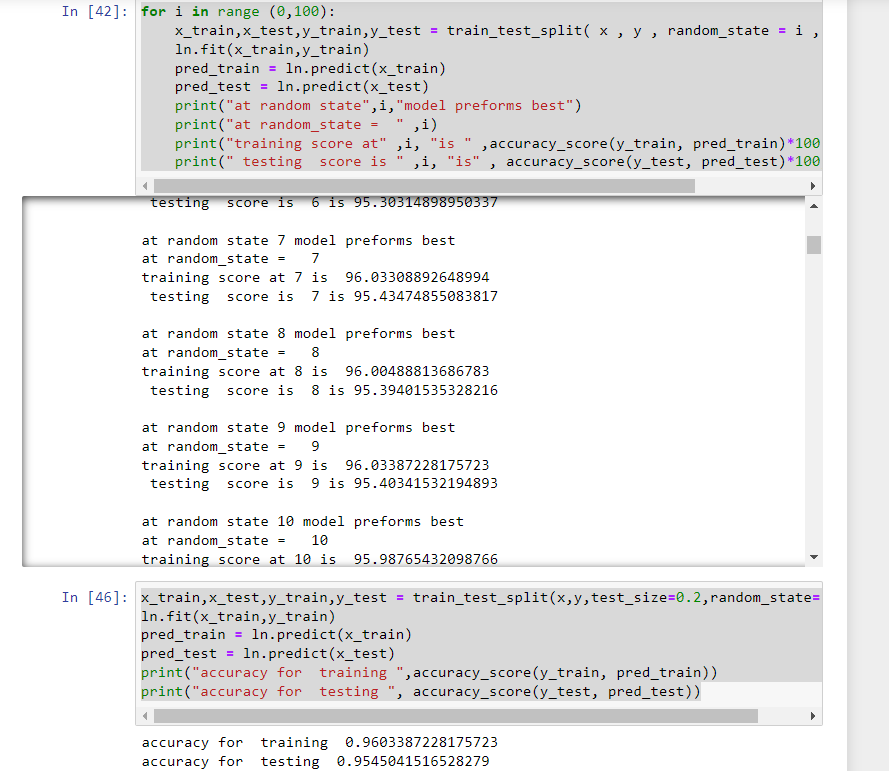
accuracy 0.95 31915

macro avg 0.94 0.78 0.84 31915

weighted avg 0.95 0.95 0.95 31915

[[28722 134]

[ 1318 1741]]



Multinomial NB gives accuracy for testing 0.9207895

973680088

Precision , recall , f3 , support as follows

precision recall f1-score support

0 0.92 1.00 0.96 28923

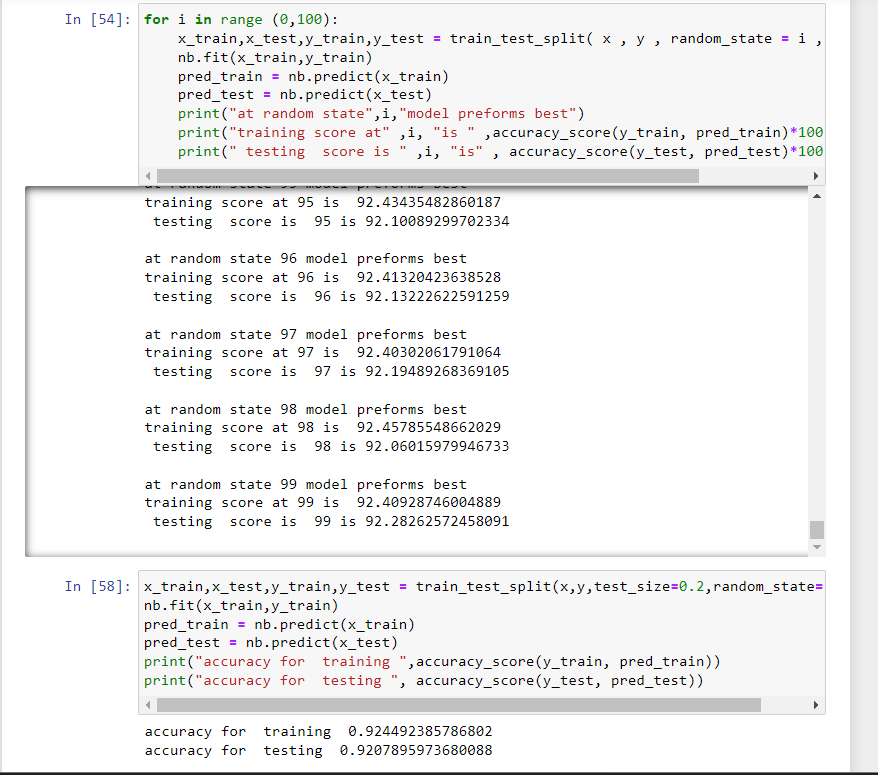
1 0.99 0.18 0.30 2992

accuracy 0.92 31915

macro avg 0.95 0.59 0.63 31915

weighted avg 0.93 0.92 0.90 31915

[[28916 7]

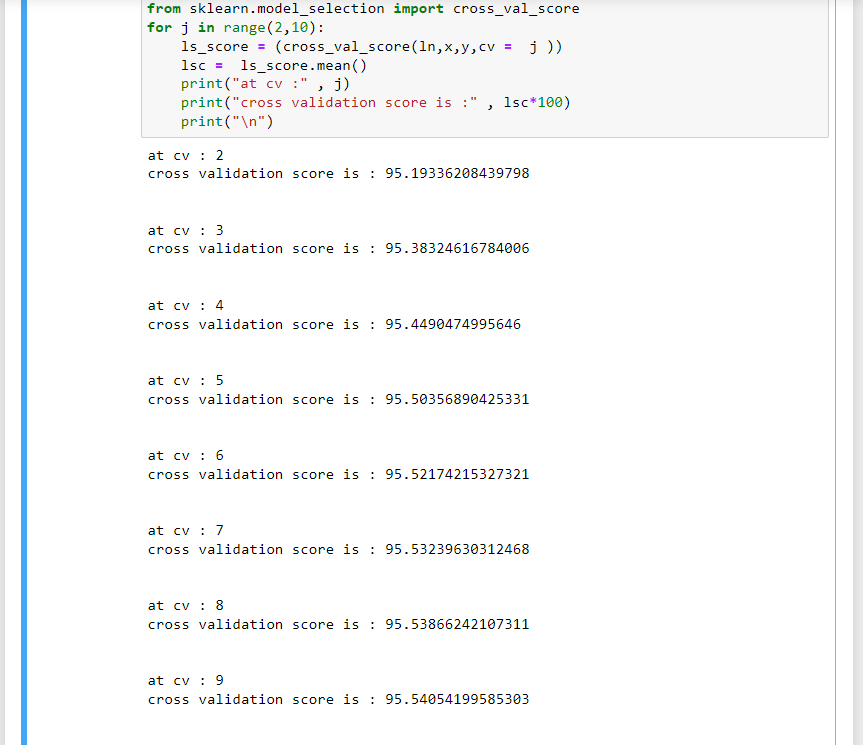
[ 2456 536]] 

**After the grid search cv the accuracy for training 0.9603387228175723**

**accuracy for testing 0.9545041516528279**

**cross validation score at 9 is the highest \ at cv : 9**

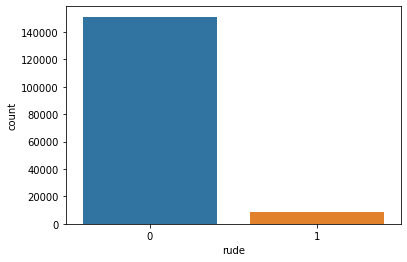
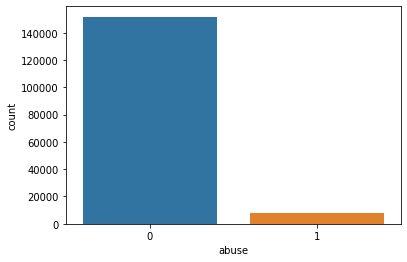
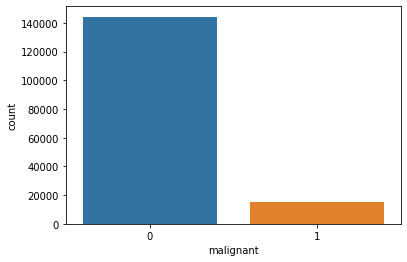
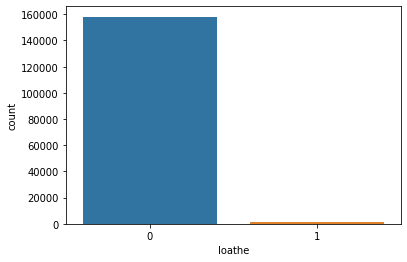
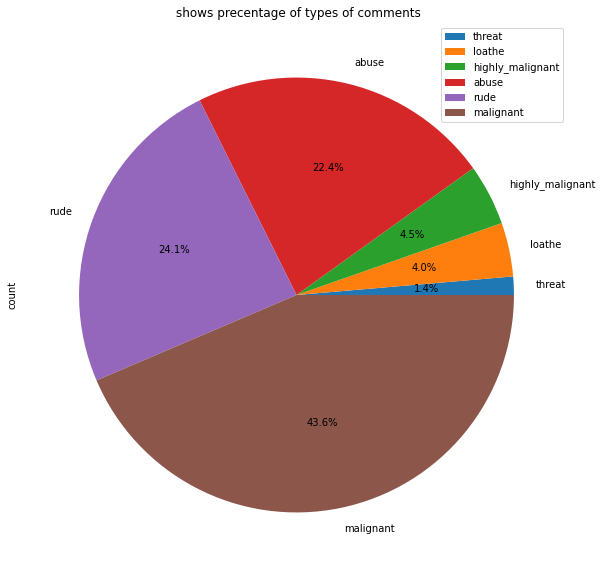
**cross validation score is : 95.54054199585303**



* Key Metrics for success in solving problem under consideration

Metrics used are : classification report and confusion matrix

**precision recall f1-score support**

* Visualizations
* **   **
* ****
* Interpretation of the Results

The final model gives accuracy of 0.95 with grid search and cv score of 95.54 at cv = 9

Roc AUC curve gives a score of 0.966

