-

MALIGNANT COMMENTS CLASSIFICATION

Submitted by:

BISHWAJIT BHATTACHARYA

ACKNOWLEDGMENT

Here target data is Integer so classification technique is used. We have total 159571 data. As the data type is mixed of string and int, taken encoder for eda process.

INTRODUCTION

Problem Statement

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.

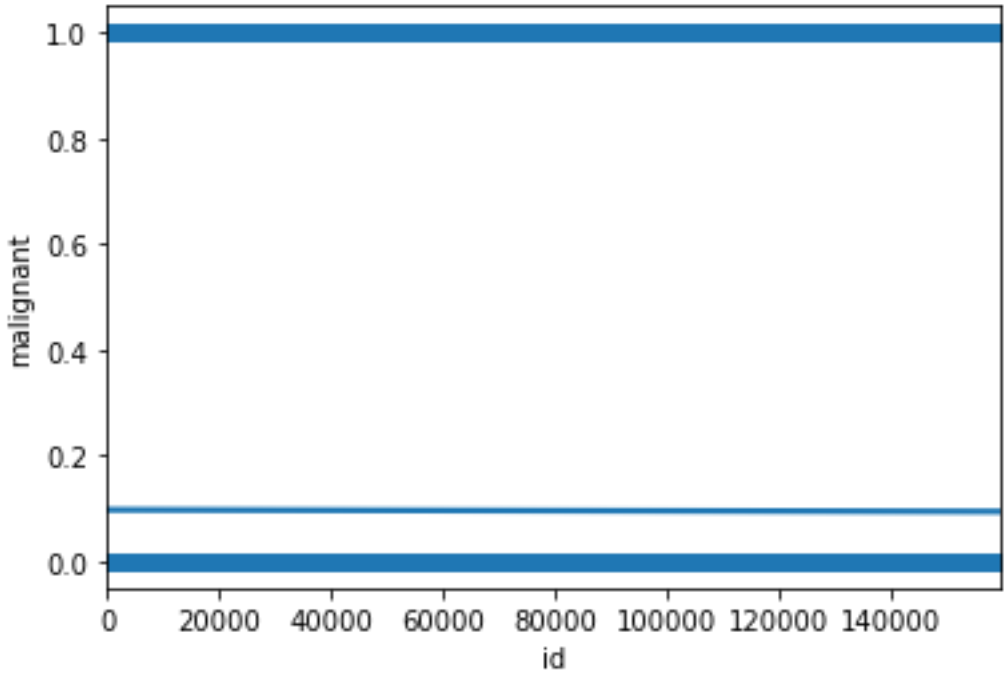
Our goal is to build a prototype of online hate and abuse comment classifier which can used to classify hate and offensive comments so that it can be controlled and restricted from spreading hatred and cyberbullying.

EDA Processing

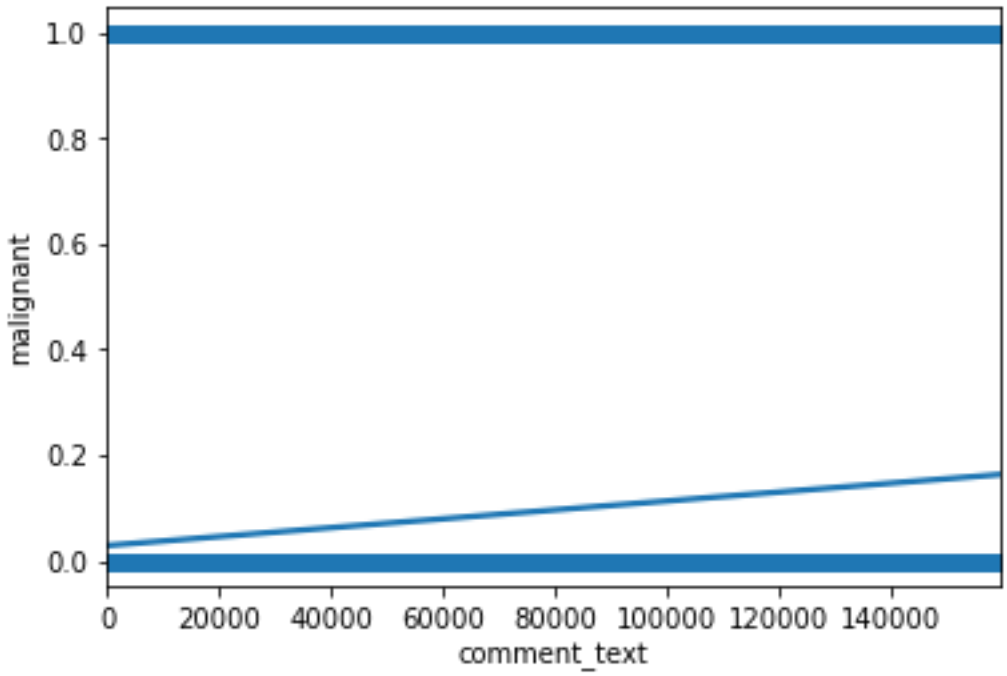
Basically, here two input variables Id and comments on text and output variables are Malignant, Highly Malignant, Rude, Threat, Abuse and loathe.

Now we will check eda for all step by step.

Id vs Malignant



Comments on Text vs Malignant

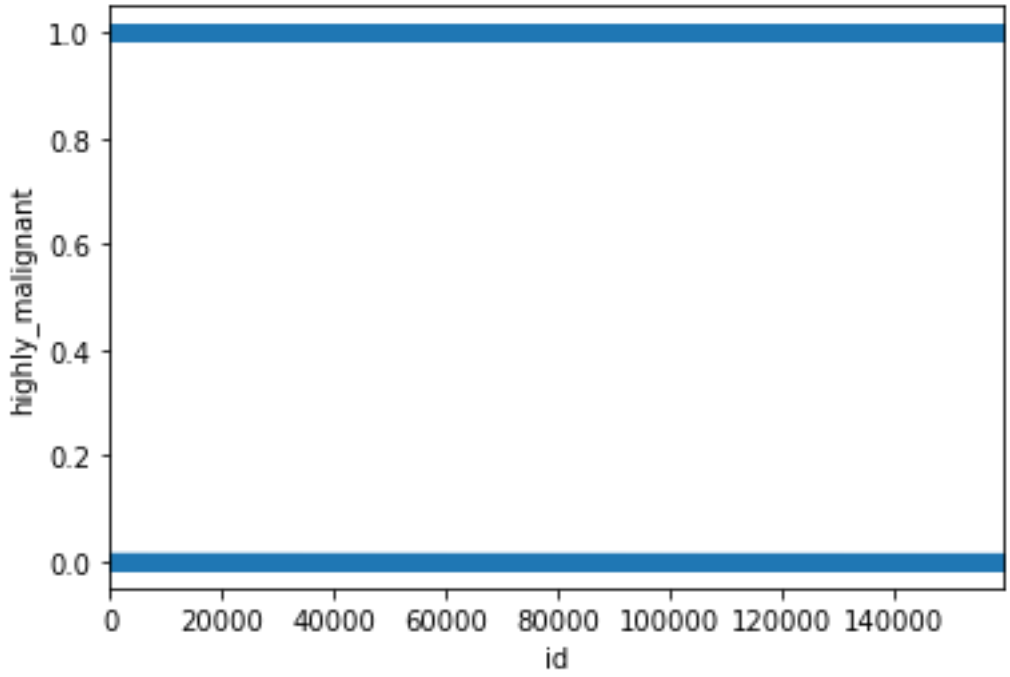


In this two graphs we can see malignant term is depended on comments on text, it taken as 0 and 1. O terms is for no malignant and 1 is when Malignant.

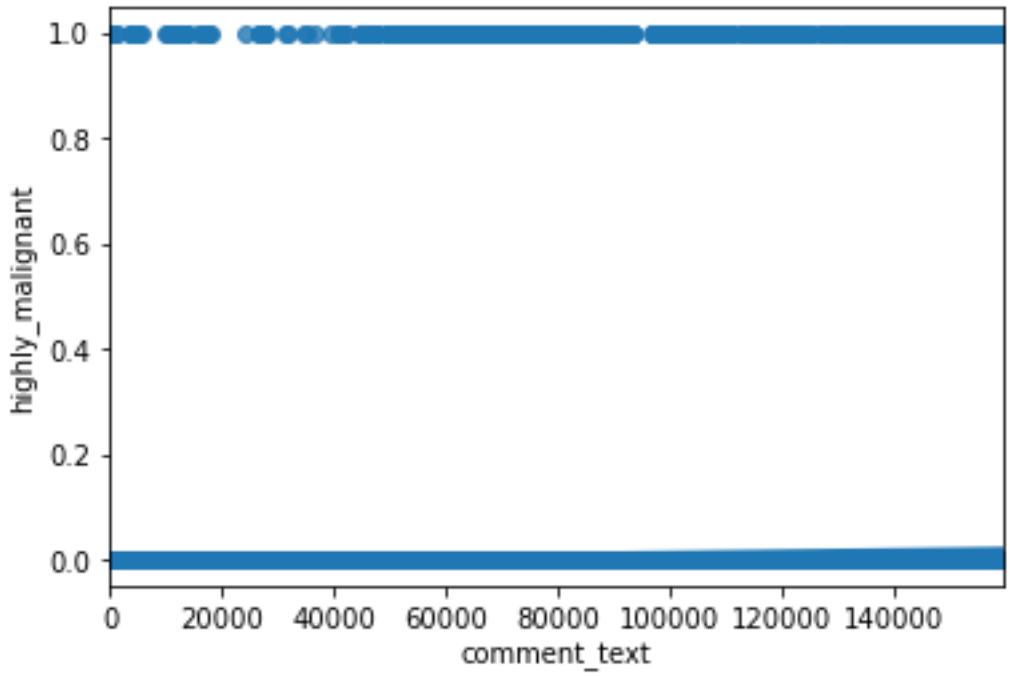
Same thing is for other out output variable as well,

Id vs Highly Malignant

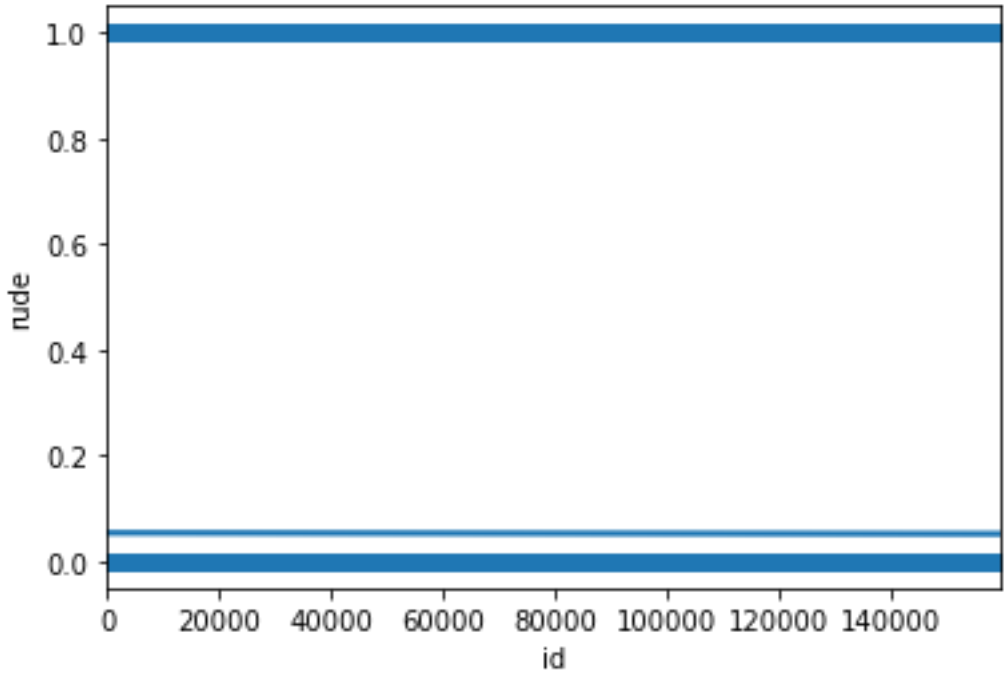
```
esSubplot:xlabel='id',
ylabel='highly_malignant'>
```



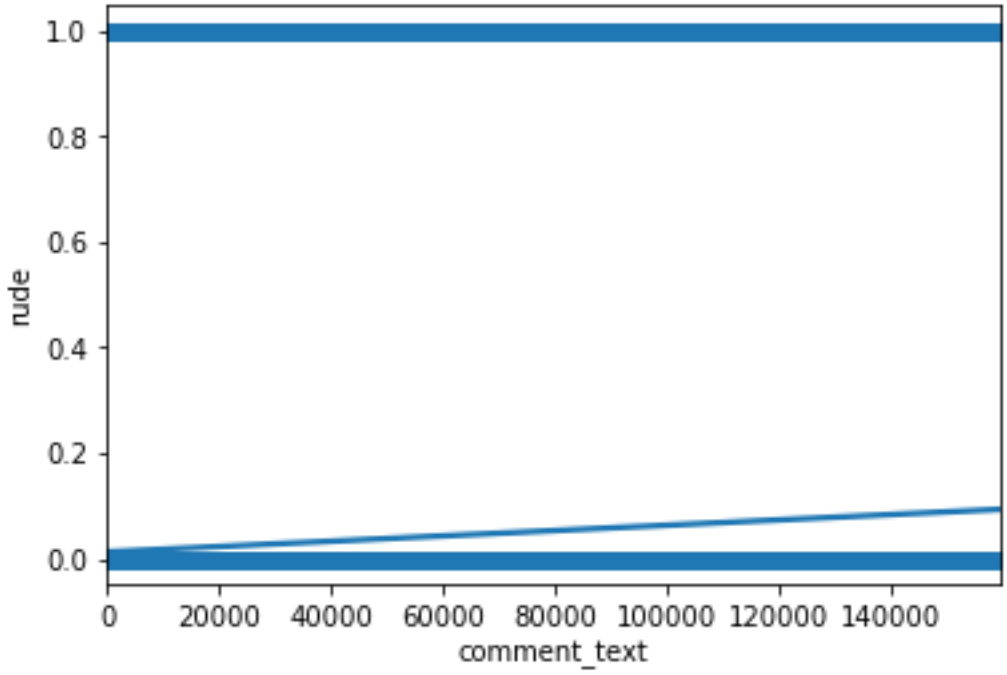
Comments on Text vs Highly Malignant



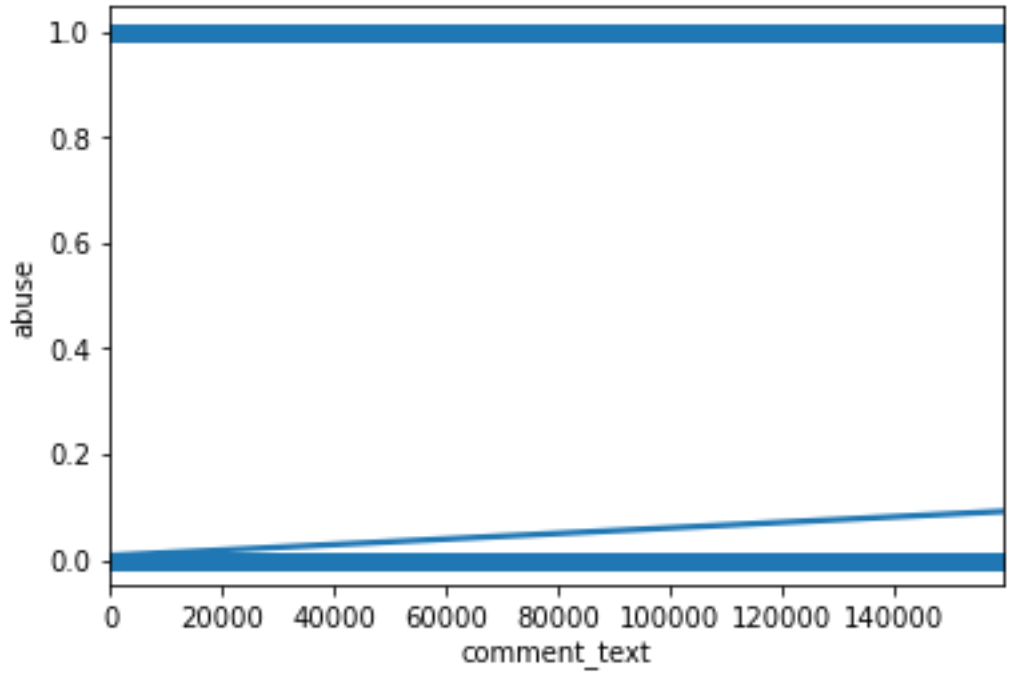
Id vs Rude



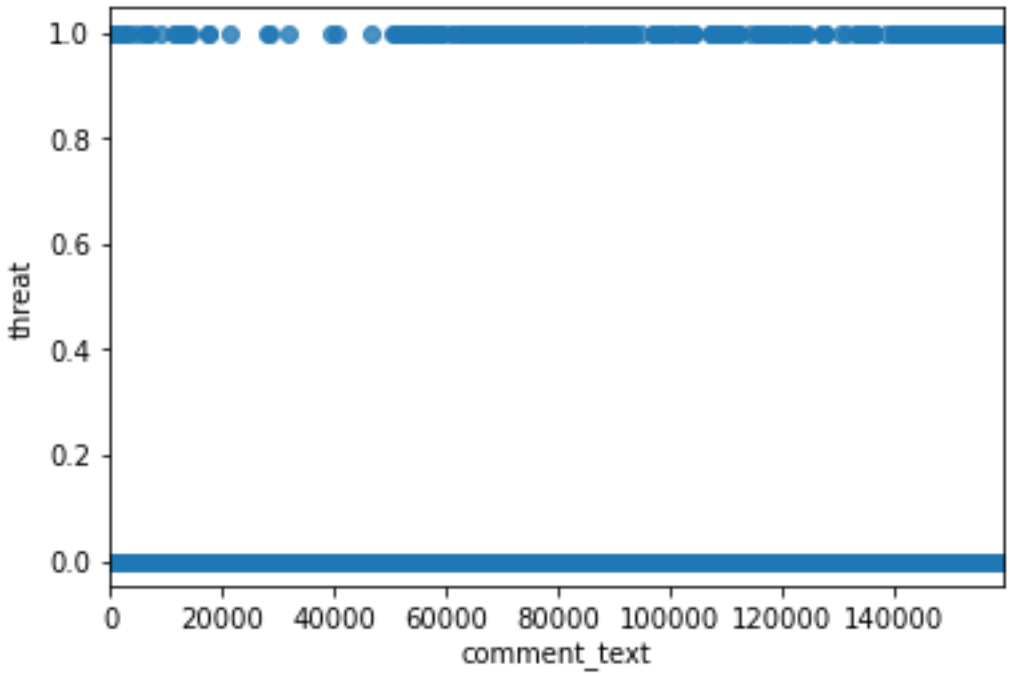
Comments on text vs Rude



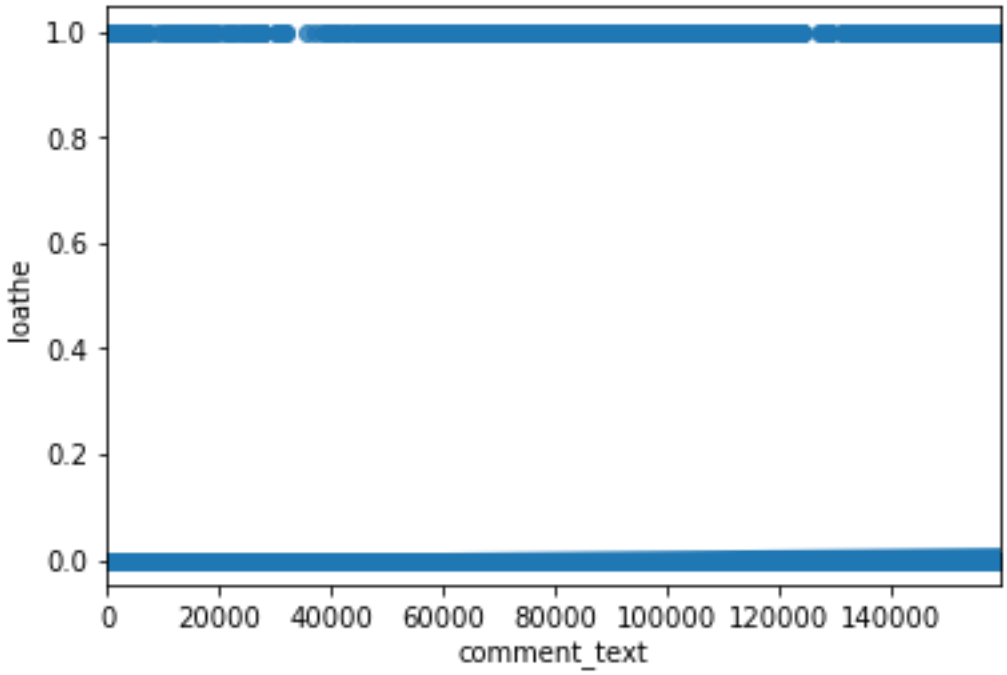
Comments on Txt vs Abuse



Comments on Txt vs Threat

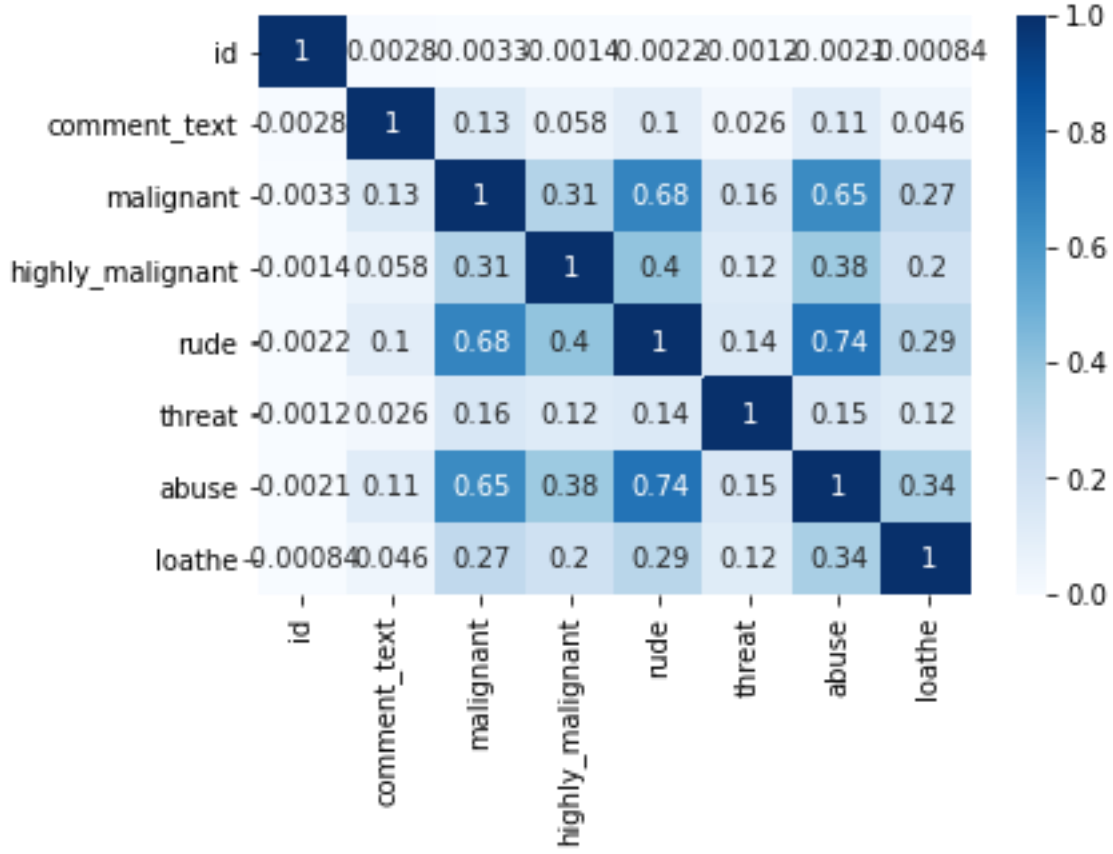


Comments on Txt Vs Loathe



Here we will see all the content with comments on text.

Heatmap



As we have 6 target value, I created 6 model separately with classification technique.

```
0.9911387405866818
```

Means 99.11% for Malignant.

```
0.9897099088257818
```

Means 98.97% for Highly_ Malignant

```
0.9642228415080397
```

Means 96.42% for Rude

```
0.9946106846310512
```

Means 99.46% for Abuse.

0.963778787404042
Means 96.38% for threat

And last
0.9907253642487859

Means 99.07% for Loathe

We use Dtc , Knn and Mnb Classifier.

Conclusion: -

Comments on any social media or any platform is very much essential. Unethical or abusive comments has a special effect on this matter. The Model which I generated it can predict almost 99% malignant comments. Also it can predict Loathe , abuse, threat , highly malignant out put.

Limitation of this programs is , as wanted to get better eda I encoded Comments on text and id starting only, so if we want to get output for different data , need to encode those two row first.

Thank You