Malignant Comments Project Report

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.
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.

Steps taken:

1. We check the data inside the dataset to understand the type of data.
2. We see that there are no null values present in the dataset.
3. Now, we look into the comment feature of the data closely.
4. There are a lot of data cleaning to be done, hence we do that.
5. After the data cleaning, we look into the data again to infer insights.
6. Now we do a general data analysis to understand the classification.
7. Now we do an extensive research using a correlation matrix to grade each column to the weightage to the target column.
8. We remove the features that has very little correlation.
9. Now we have reduced our dataset to the features that produces a weightage to the target column and hence ready to train the model.
10. We test different models to short list the best performing one.

Conclusion:

From the project we get a better understanding of the toxic words that are included in the sentence that make the comment malicious. On this basis and understanding, we can easily infer that sentences containing such words and on the frequency of the appearance of these words classifies the comment as rude, threat, loathe or abuse.