

Malignant Comments Classifier

Submitted by:

RAVI VERMA

**ACKNOWLEDGMENT**

It is indeed a great pleasure and matter of immense satisfaction for us to express our deep sense profound gratitude towards all the people who have helped, inspired me in my project work.

I, would like to thanks **Flip Robo Technologies** which was always there whenever we needed any support and was a constant source of inspiration for accomplishment of this project.

For the following project I used certain references which are mentioned below:

1. towardsdatascience.com
2. medium.com
3. stackoverflow.com
4. bing.com

**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.

* Review of Literature

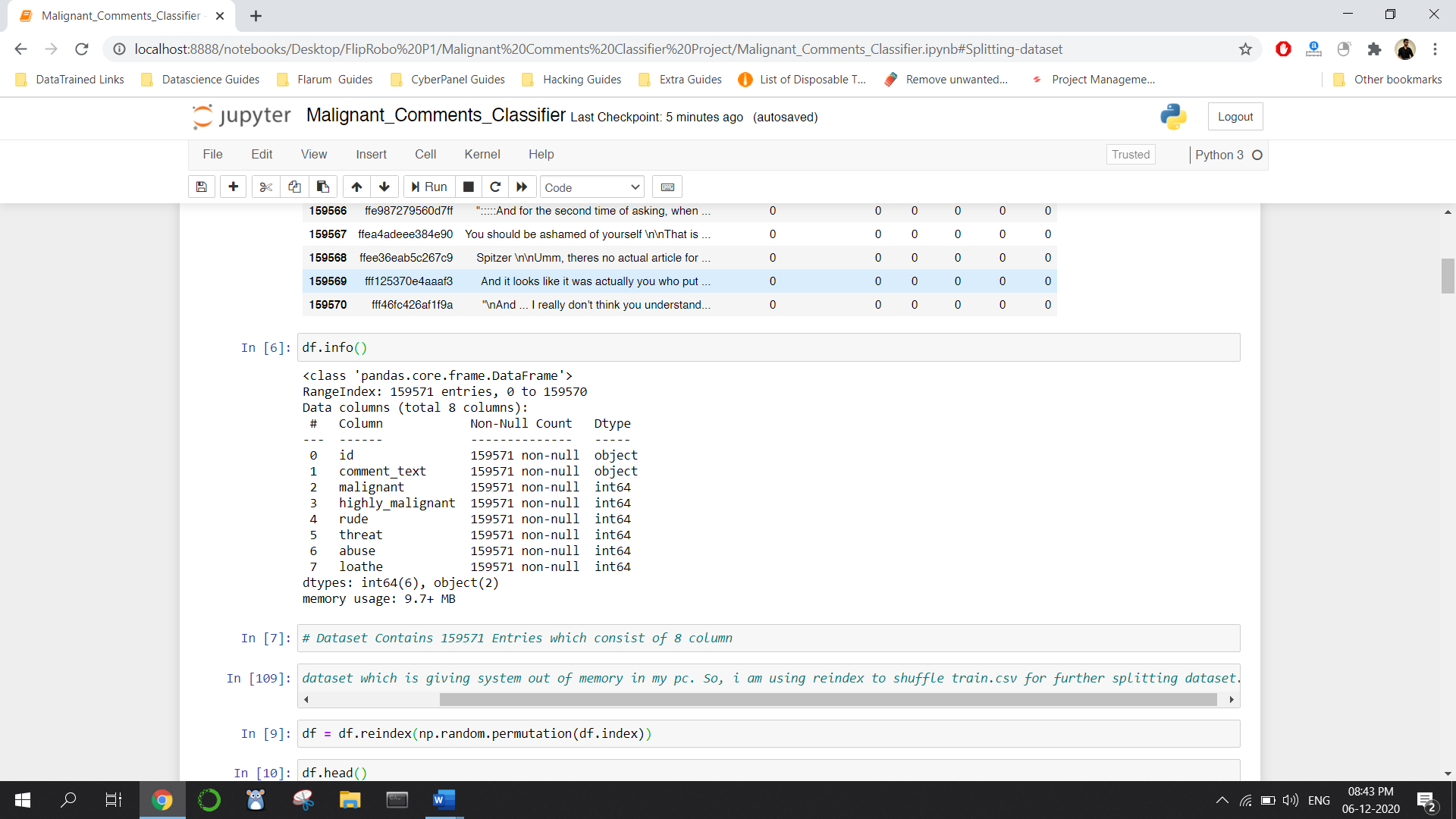
While we were working on this project we need to go through many sources, books and references. We are sharing some of the content which is relevant and useful in the lieu of this particular project.

* import pandas as pd
* import numpy as np
* import matplotlib.pyplot as plt
* import matplotlib.cm as cm
* import itertools
* import string
* from stop\_words import get\_stop\_words
* import nltk
* from nltk.stem import PorterStemmer, WordNetLemmatizer
* from sklearn.feature\_extraction.text import CountVectorizer
* from sklearn.metrics import hamming\_loss
* from sklearn.metrics import accuracy\_score
* from sklearn.metrics import log\_loss
* from sklearn.naive\_bayes import MultinomialNB
* from skmultilearn.problem\_transform import BinaryRelevance
* from sklearn.svm import SVC
* from sklearn.naive\_bayes import GaussianNB
* from skmultilearn.problem\_transform import ClassifierChain
* from skmultilearn.problem\_transform import LabelPowerset
* Motivation for the Problem Undertaken

As we can see this is a highly motivated project. As this is the real time problem of malignant comments which is getting bigger, a single malignant comment can ruin many things. This can take a toll on anyone and affect them mentally leading to depression, mental illness, self-hatred and suicidal thoughts.

**Analytical Problem Framing**

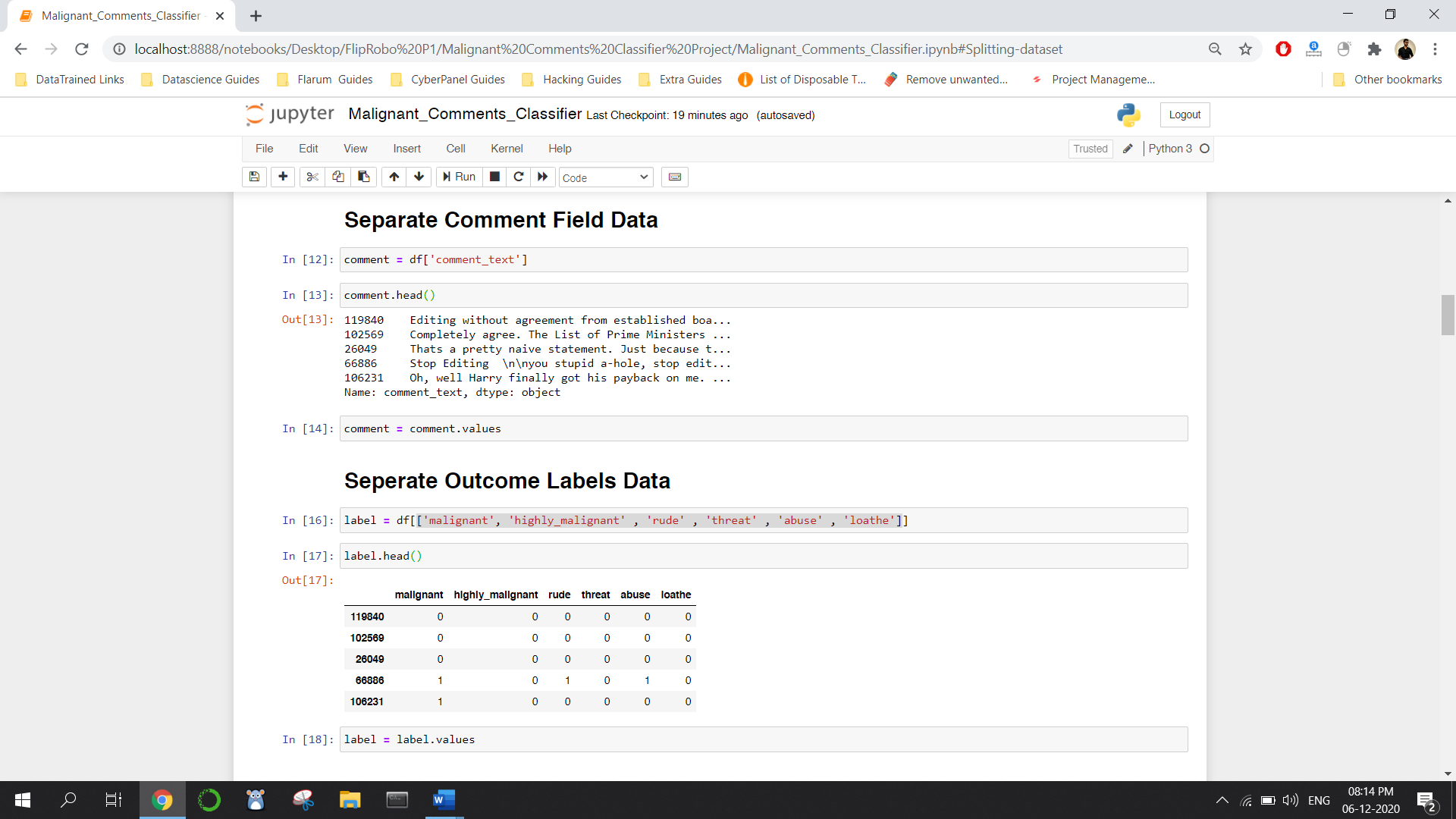
* Mathematical/ Analytical Modelling of the Problem

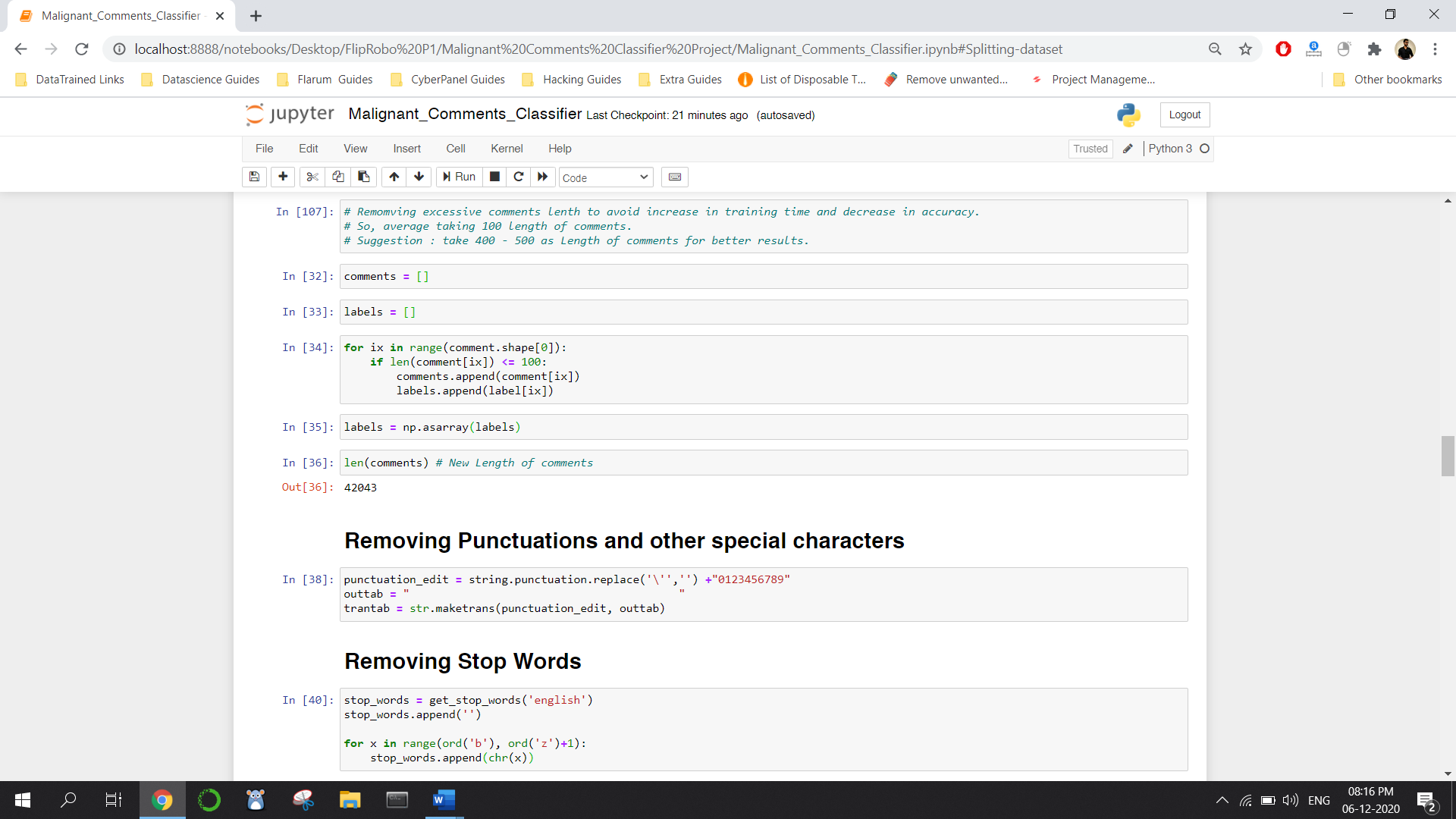


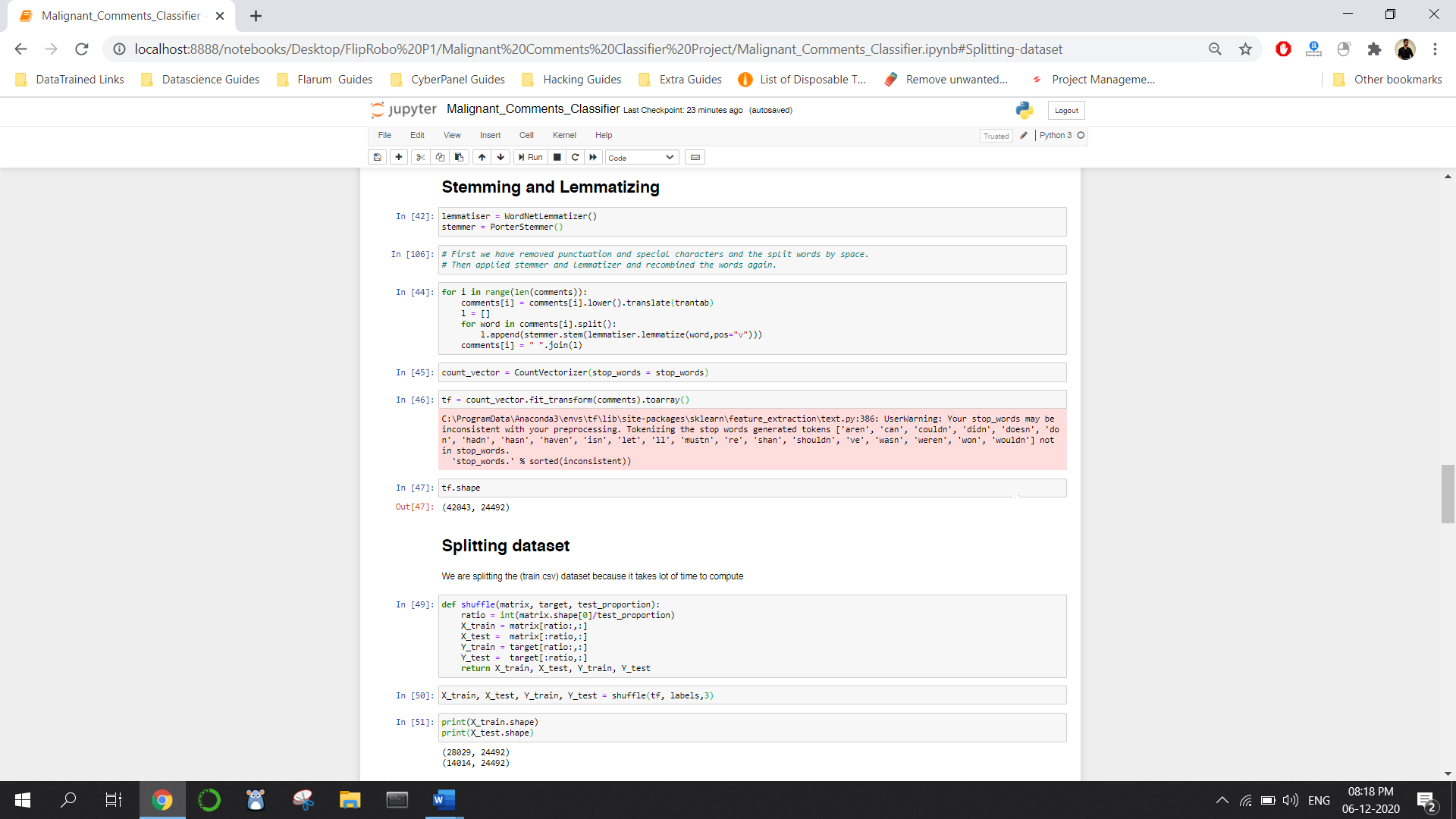
* Data Sources and their formats

There are 8 columns in the dataset provided to you. The description of each of the column is given below:

* **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.
* Data Pre-processing Done



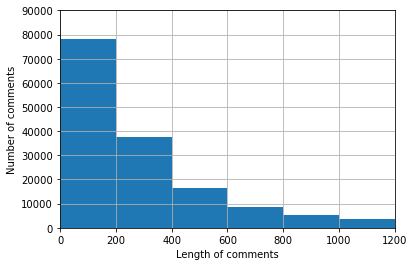


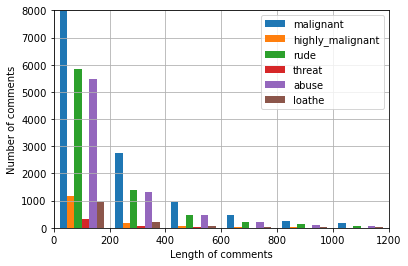


* Hardware and Software Requirements and Tools Used
* **Hardware:** A good configuration Laptop
* **Software:** Python (Jupyter Notebook)
* **Library Used:** Mentioned Above

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)





* Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

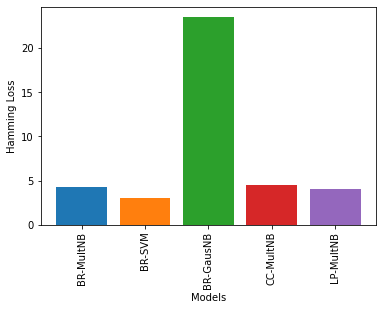
* Run and Evaluate selected models

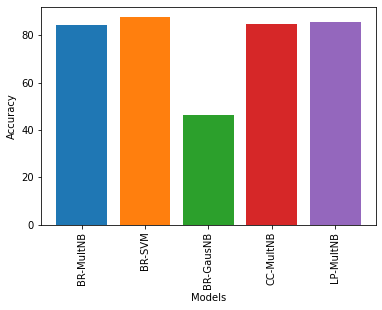
Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

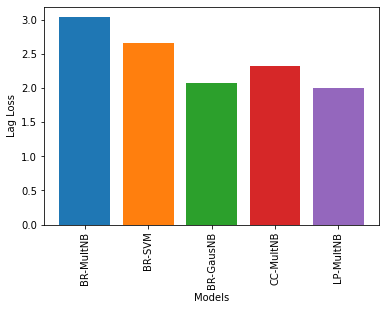
* Key Metrics for success in solving problem under consideration

What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

* Visualizations







* Interpretation of the Results

The best models work on this dataset is Label Powerset with MultinomialNB which has a Hamming Loss of 4.0, Accuracy of 85.47, Lag Loss of 2.0 and Binary Relevance with SVM which has a Hamming Loss of 3.02, Accuracy of 87.57 and Lag Loss of 2.66

**CONCLUSION**

* Key Findings and Conclusions of the Study

In the above dataset every features of the dataset plays an important role to understand the data also in visualization and applying models and algorithms.

* Learning Outcomes of the Study in respect of Data Science

Visualization is very helpful and it plays a vital role to understand the data into graphical form. So, that we can understand what the data is trying to say.

Data cleaning is also import part it helps me to clear the comments by removing stop words, punctuations, special characters, stemming and lemmatizing.

I used five types of algorithm like Binary Relevance with MultinomialNB, Binary Relevance with SVM classifier, Binary Relevance with GausseanNB, Classifier chain with MultinomialNB and Label Powerset with MultinomialNB.

The problem I faced while on this dataset is it takes more than an hour for output. I overcome this situation with the help of towardsdatascience.com by running the code using small chunk of dataset to complete the project.