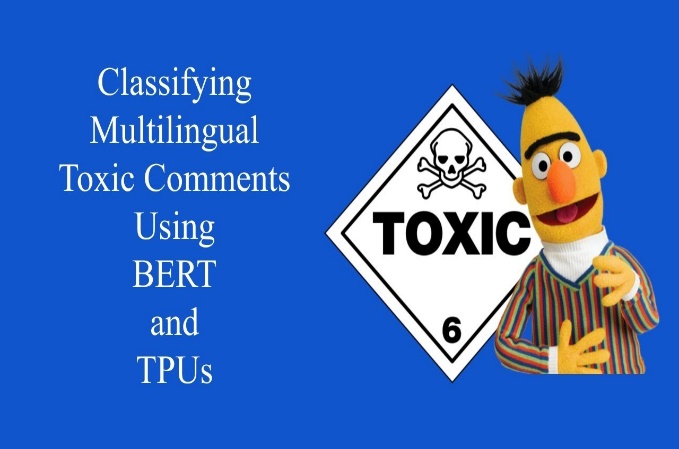
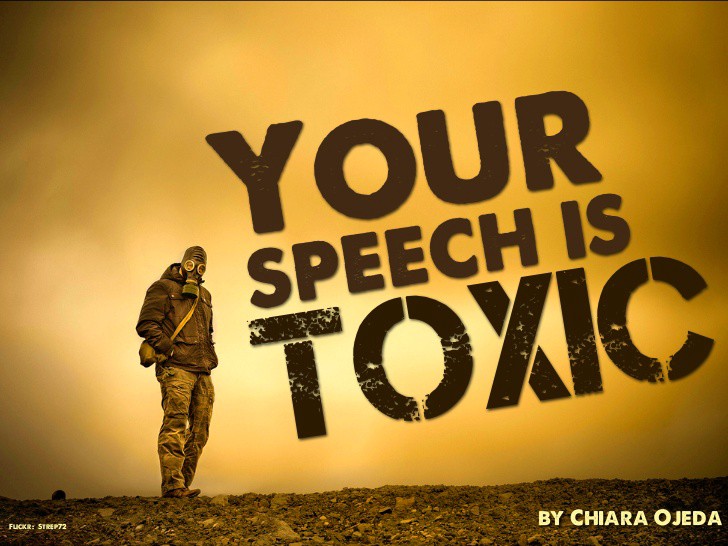


**MALIGNANT COMMENTS CLASSIFICATION**



Submitted by:

MANOJ.I.V

**INTRODUCTION**

* Business Problem Framing

Describe the business problem and how this problem can be related to the real world.

Answer: 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.

It shows for an event or opinion how would be the reaction of the people.

* Conceptual Background of the Domain Problem

Describe the domain related concepts that you think will be useful for better understanding of the project.

Answer: Frist of all we should know how to collect data by webscraping and natural language programming.

* Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what is the motivation behind.

Answer: Researchers have found that hate is a problem across multiple platforms, as there is a lack of models for online hate detection. It shows for an event or opinion how would be the reaction of the people.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

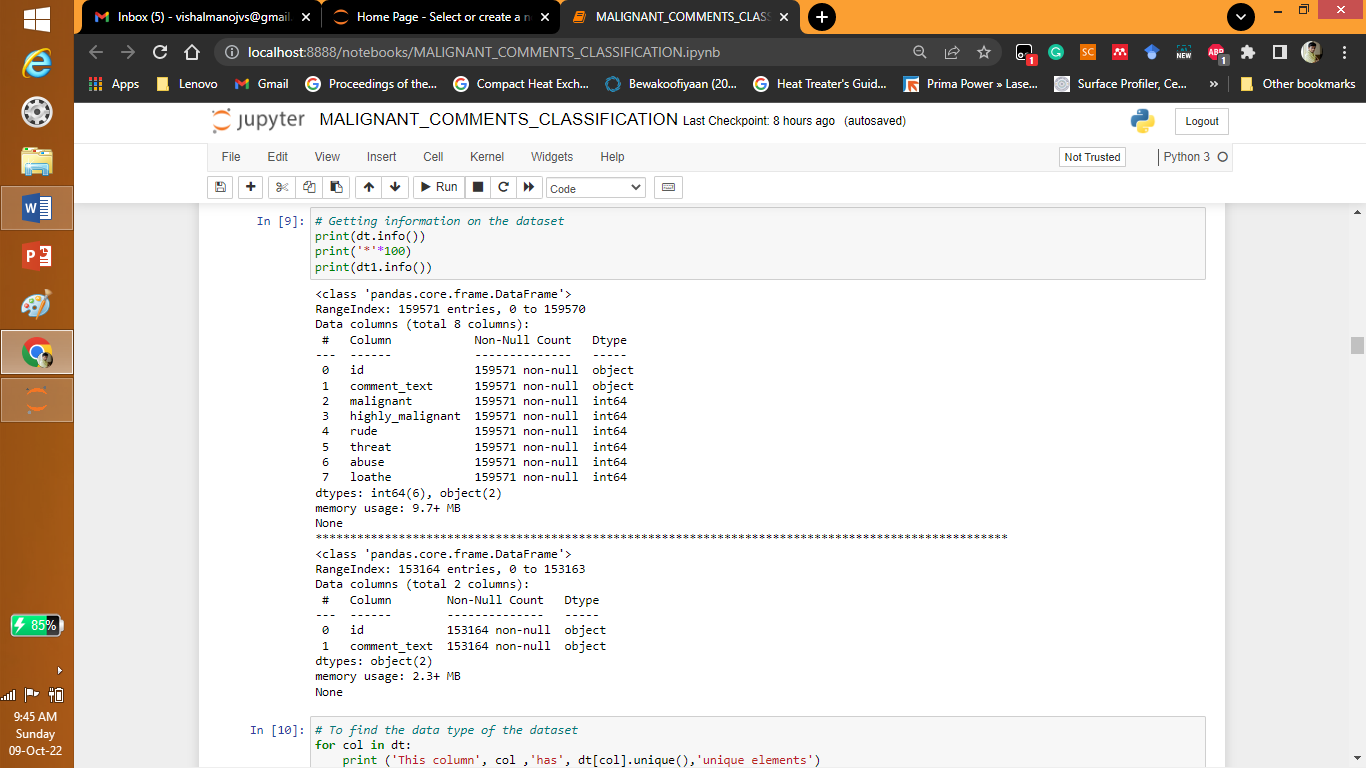
Answer: The different machine learning models used for analysis of multi-classification malignant comments are Logistic Regression, Random Forest Classifier, Support Vector Classifier, Ada Boost Classifier etc.

* Data Preprocessing Done

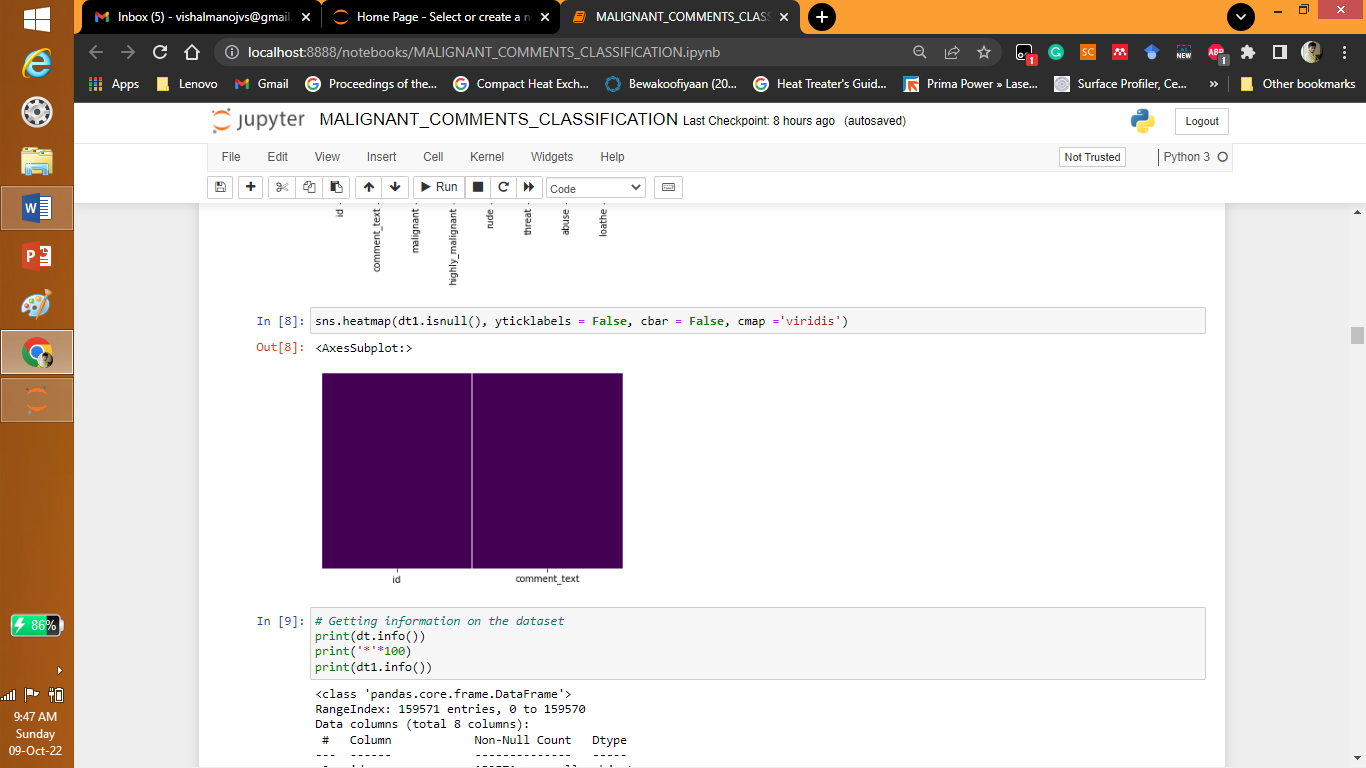
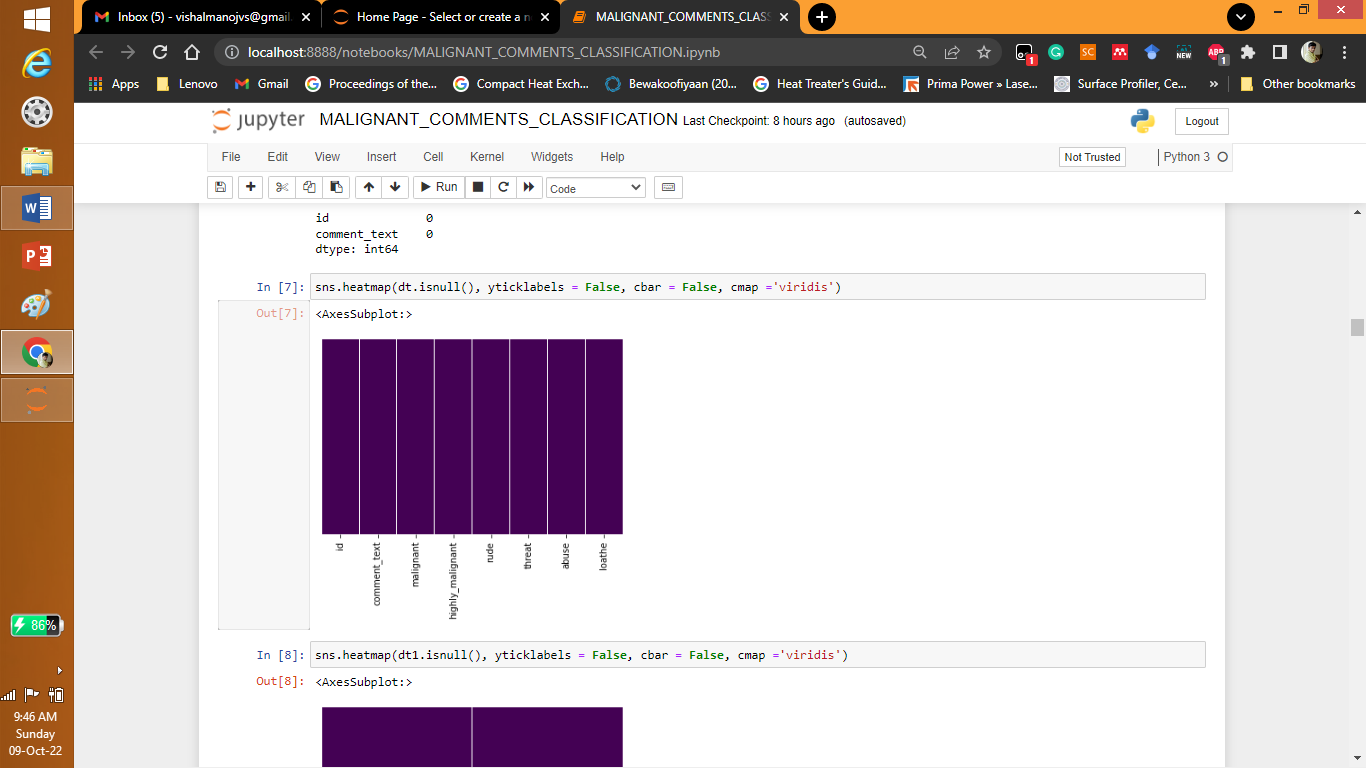
What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

Answer: The different steps followed for cleaning of data were:

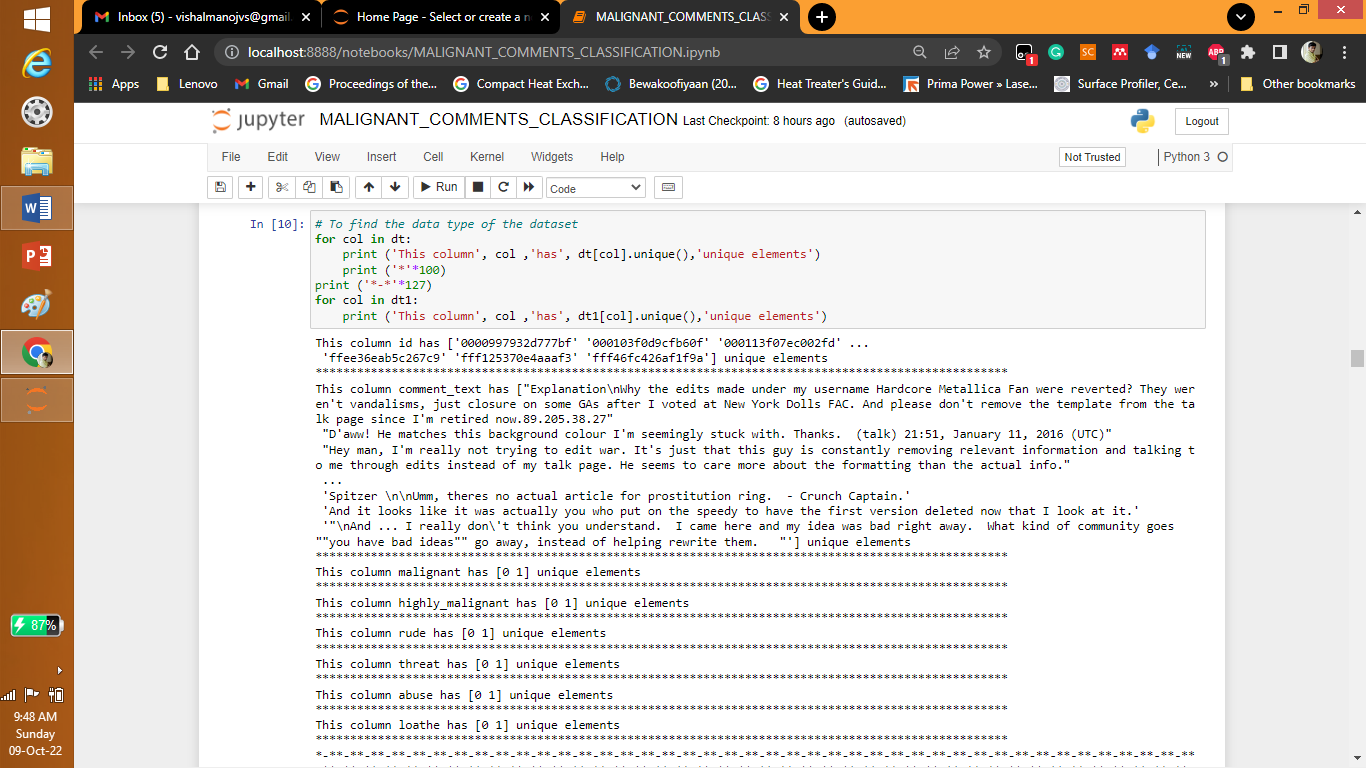
1. To check the datatype

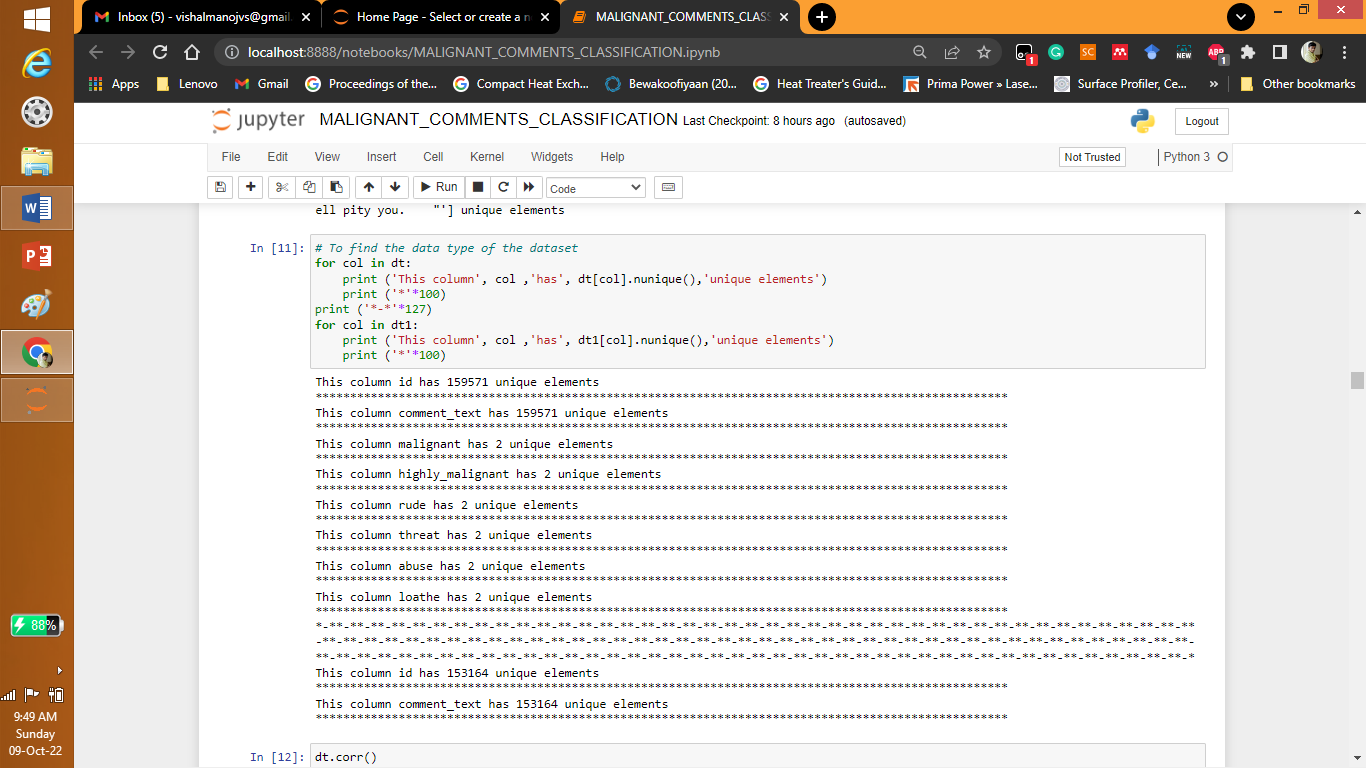


2. To check is null values



3. To check the elements

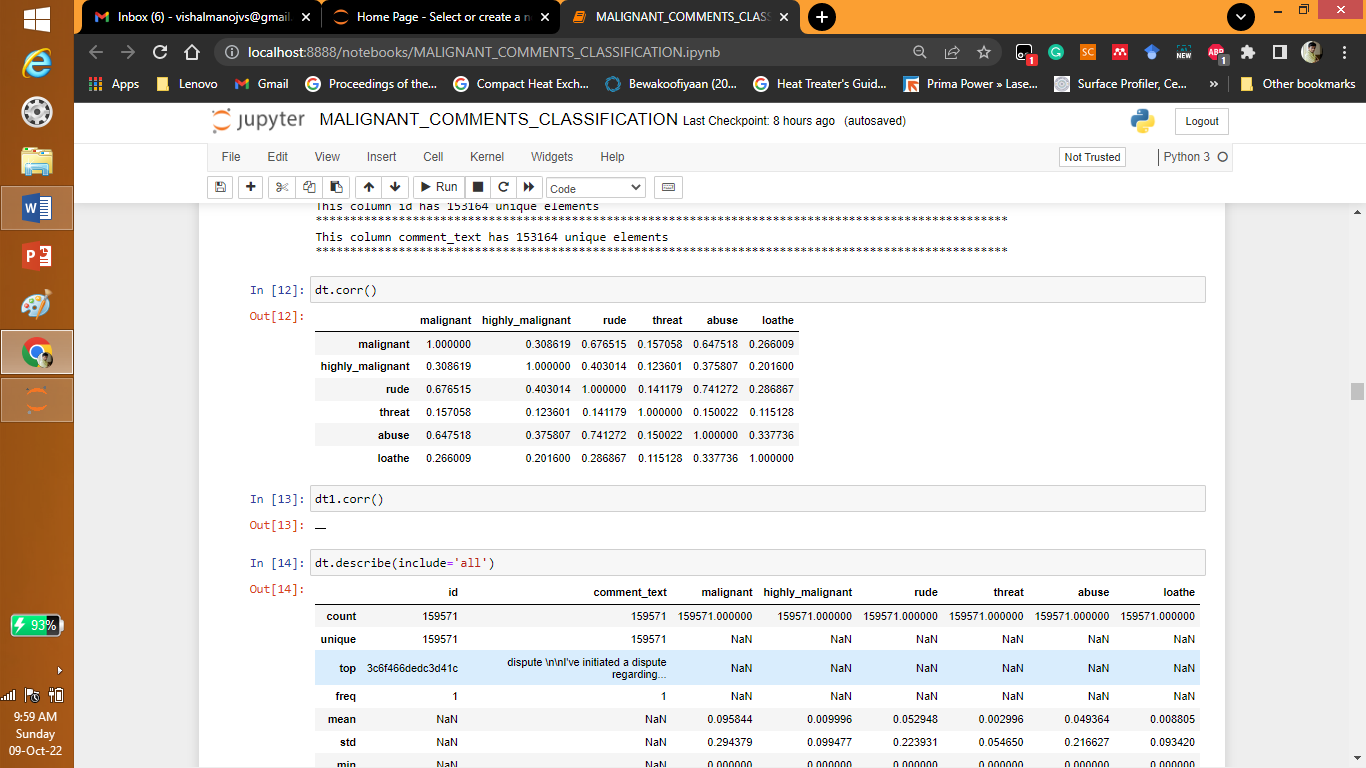




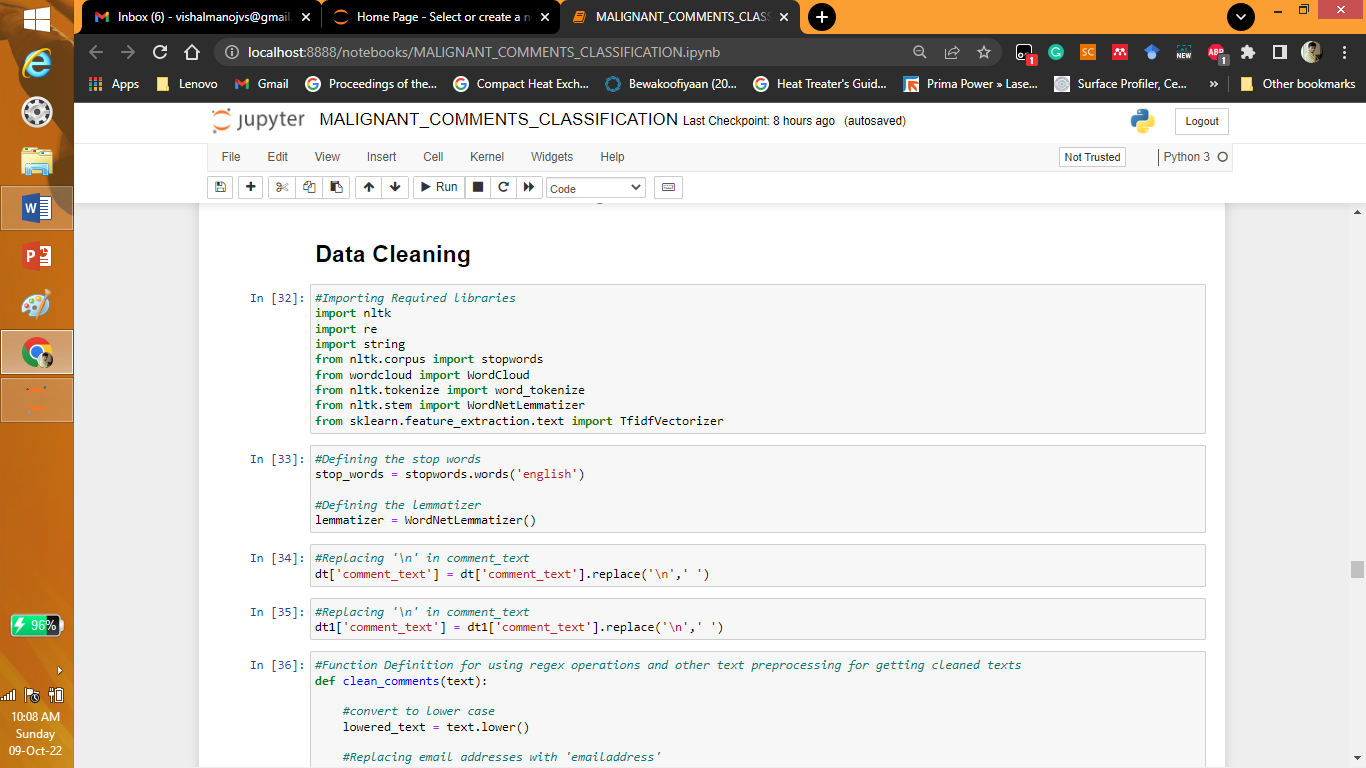
* Data Inputs- Logic- Output Relationships

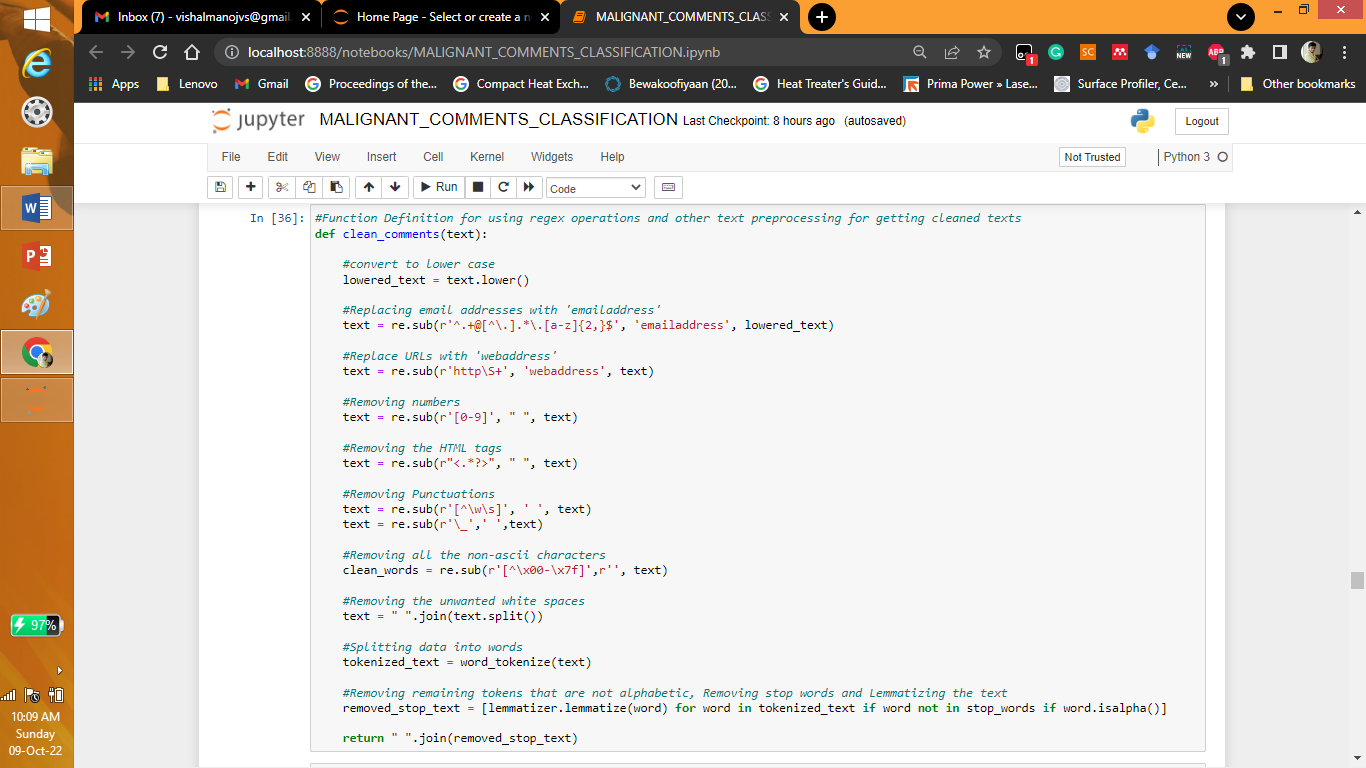
Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

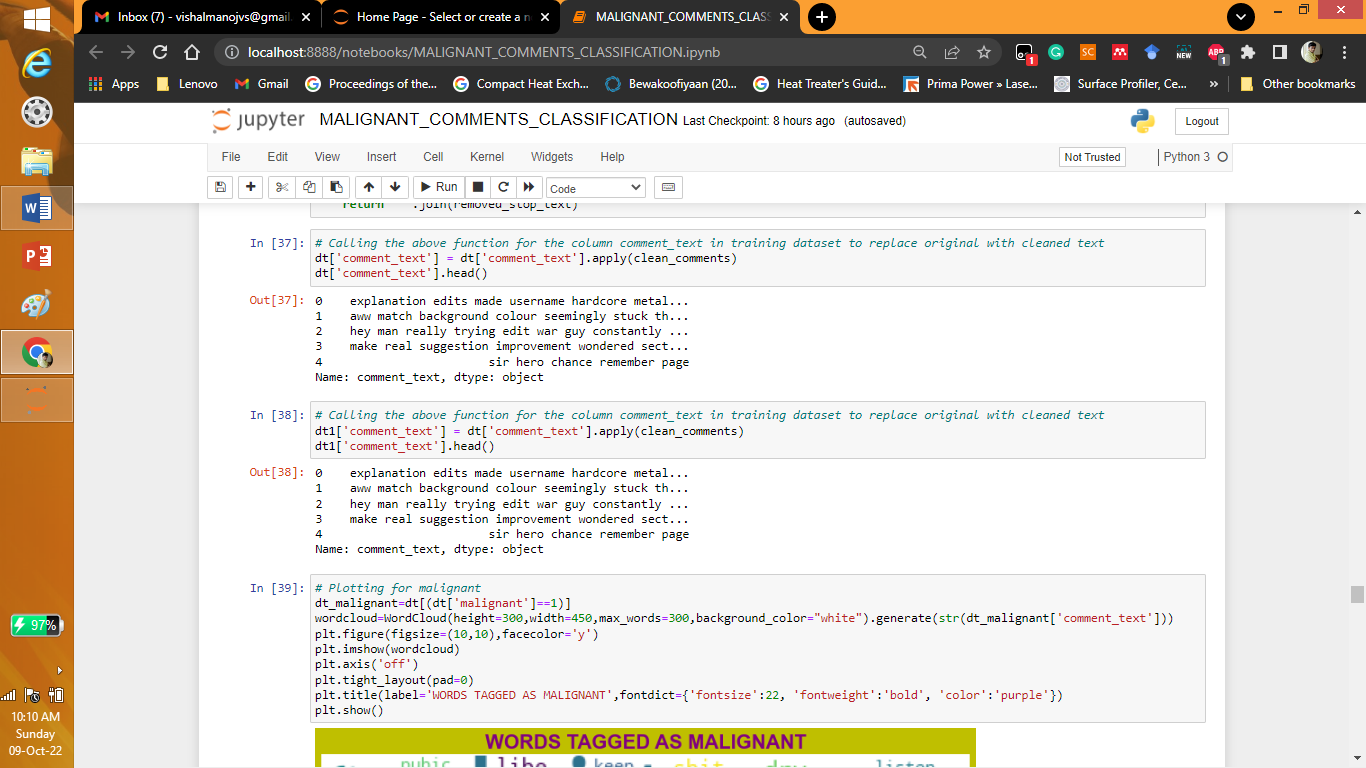
Answer: The relationship is given by











* Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

Answer: Different soft wares are Python have libraries like numpy, pandas, scikitlearn, matplotlib, seaborn, nltk, re, wordcolud etc.

**Model/s Development and Evaluation**

* Testing of Identified Approaches (Algorithms)

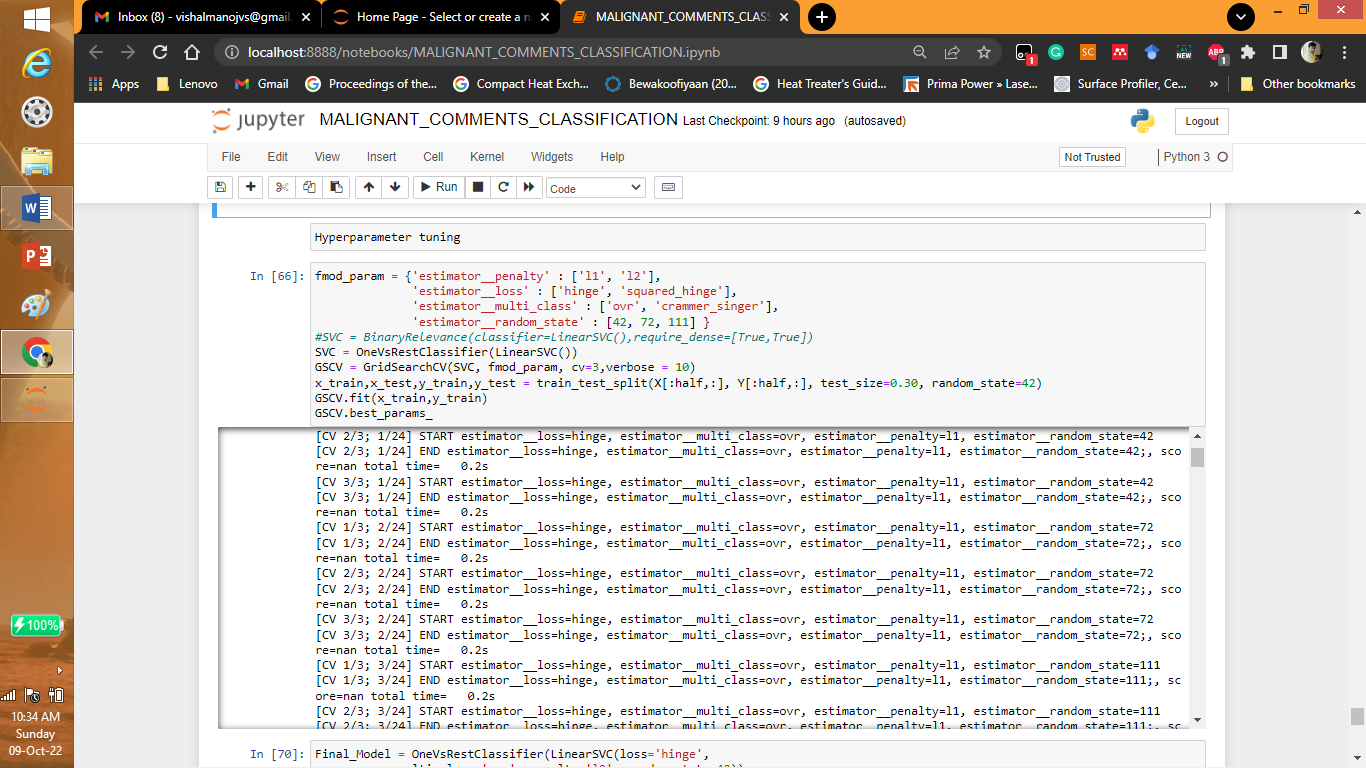
Listing down all the algorithms used for the training and testing.

Answer: The different machine learning models used for analysis of multi-classification malignant comments are Logistic Regression, Random Forest Classifier, Support Vector Classifier, Ada Boost Classifier etc

* 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.

Answer:



* 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.

Answer: Different metrics used in the problem are

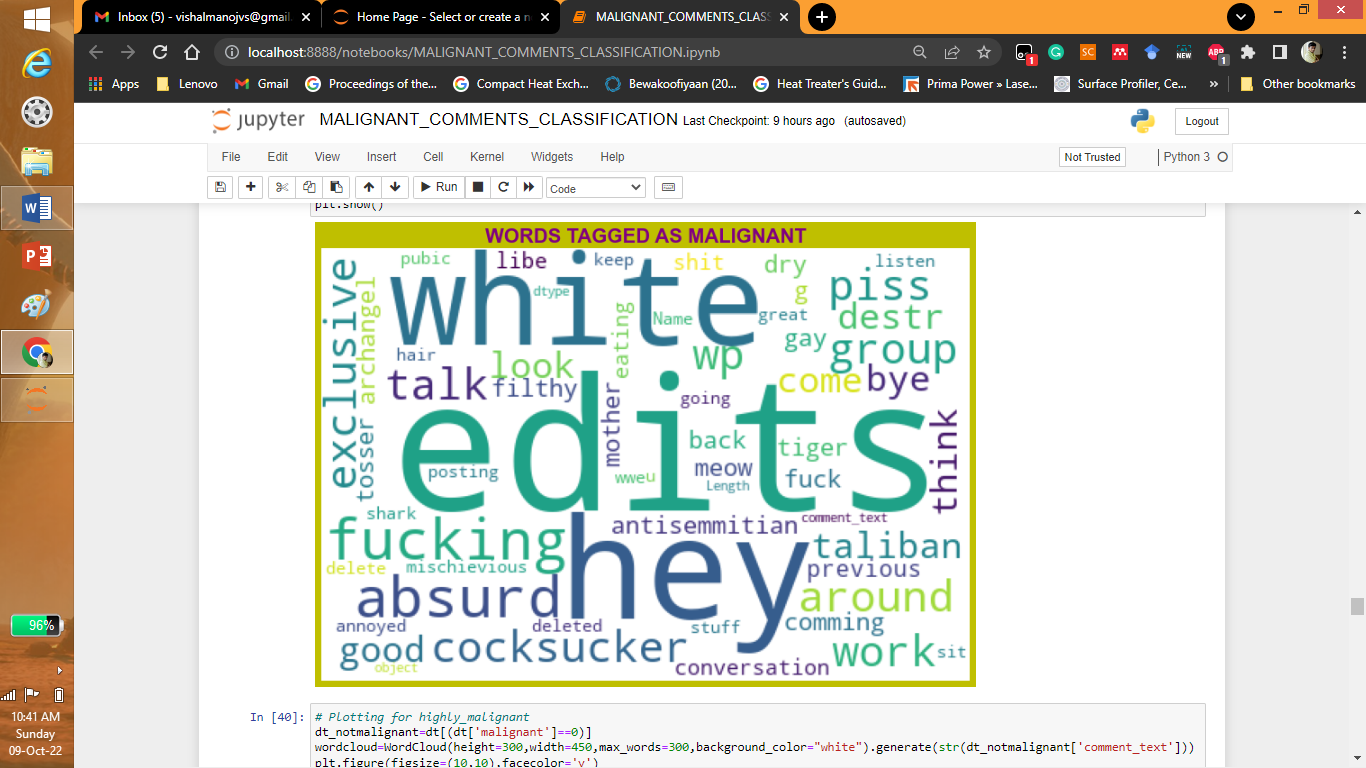
confusion\_matrix,classification\_report,accuracy\_score, hamming\_loss, log\_loss.

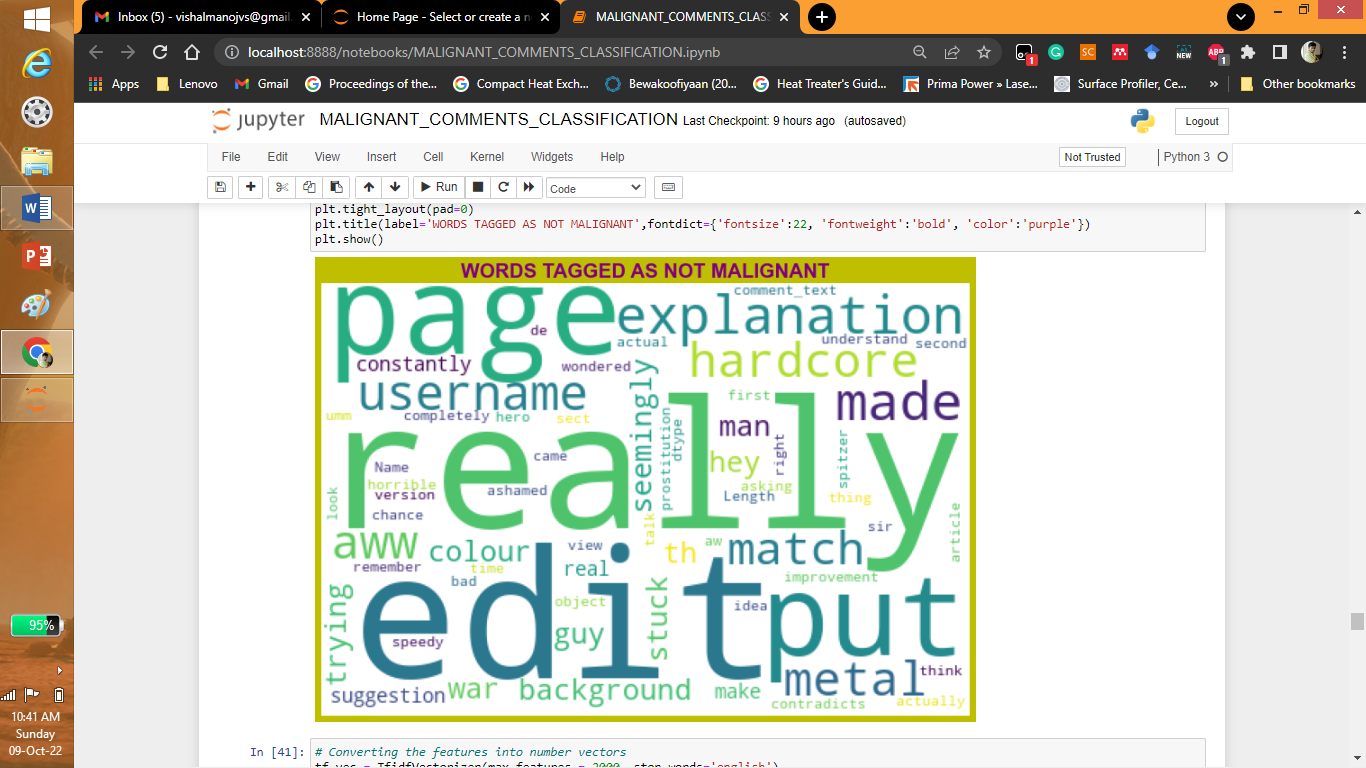
* Visualizations

Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

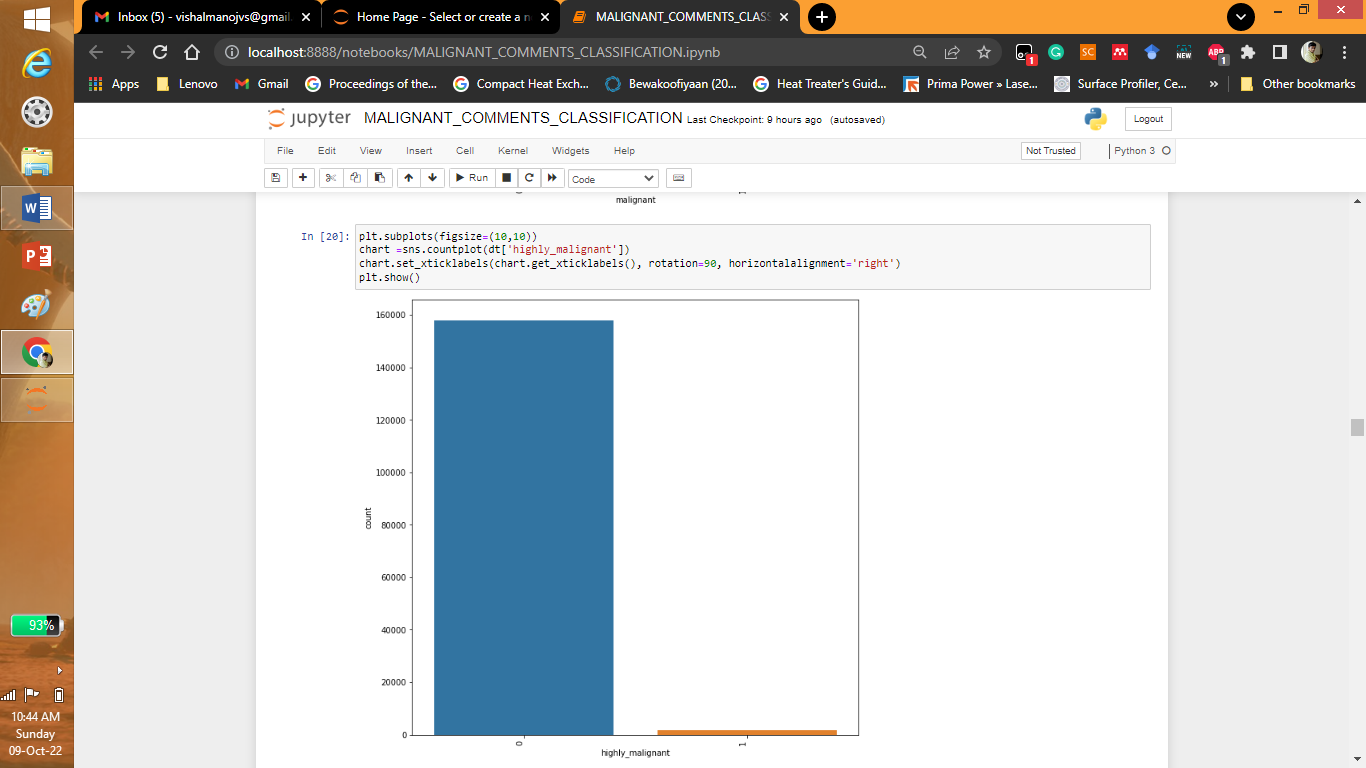
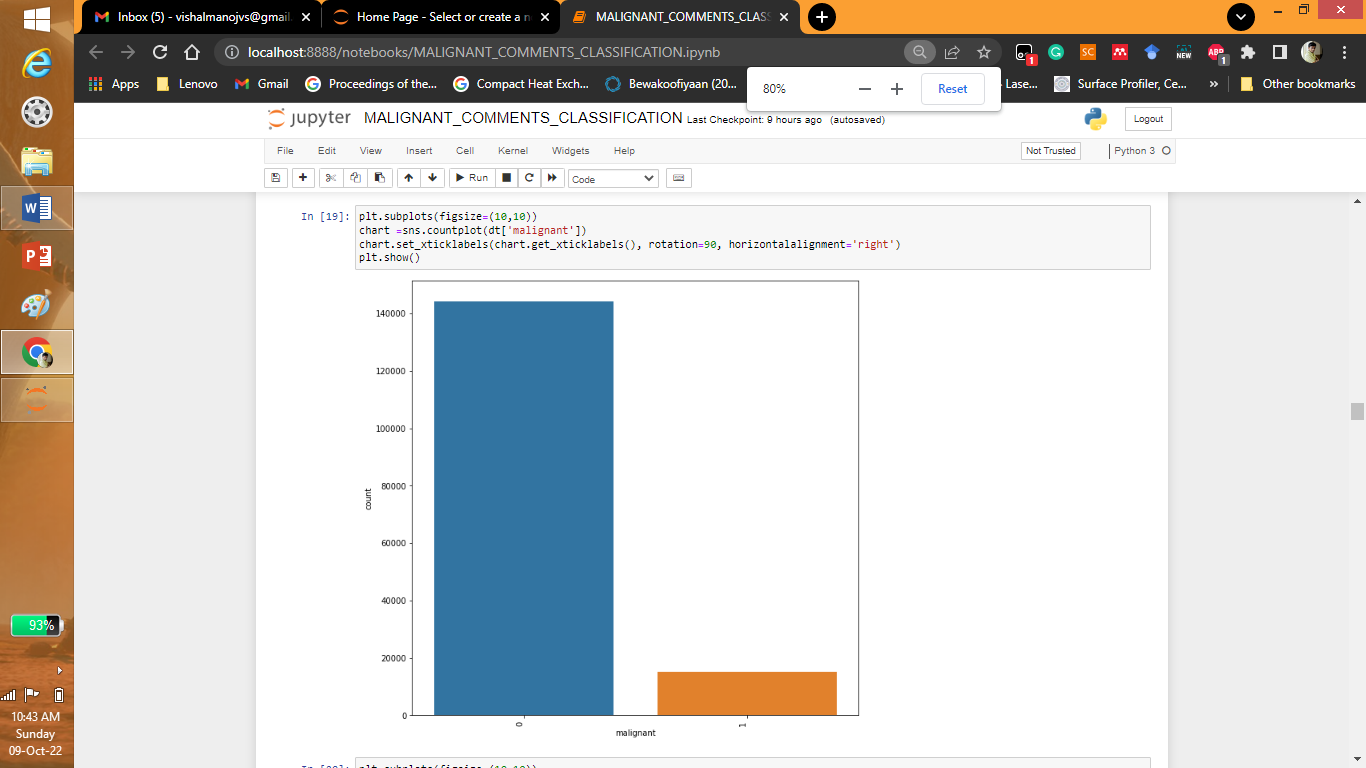
If different platforms were used, mention that as well.

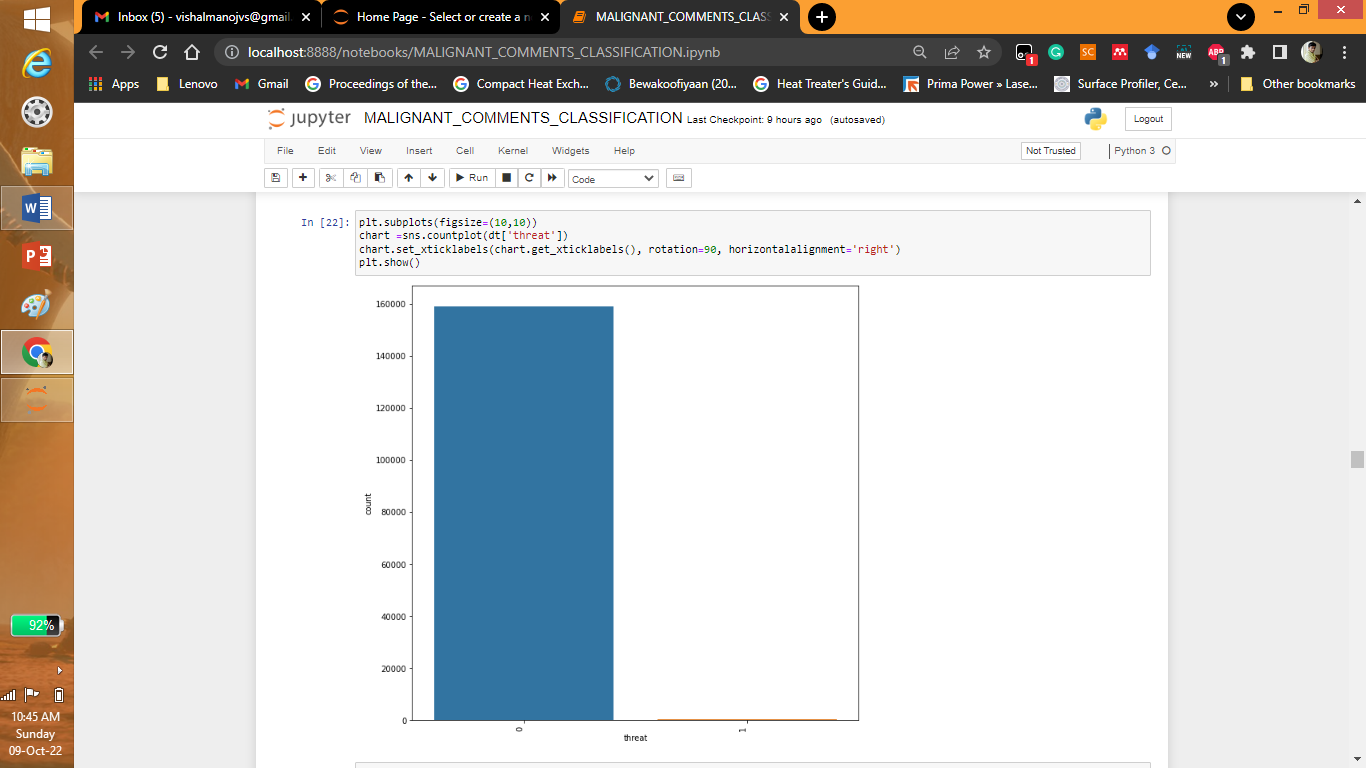
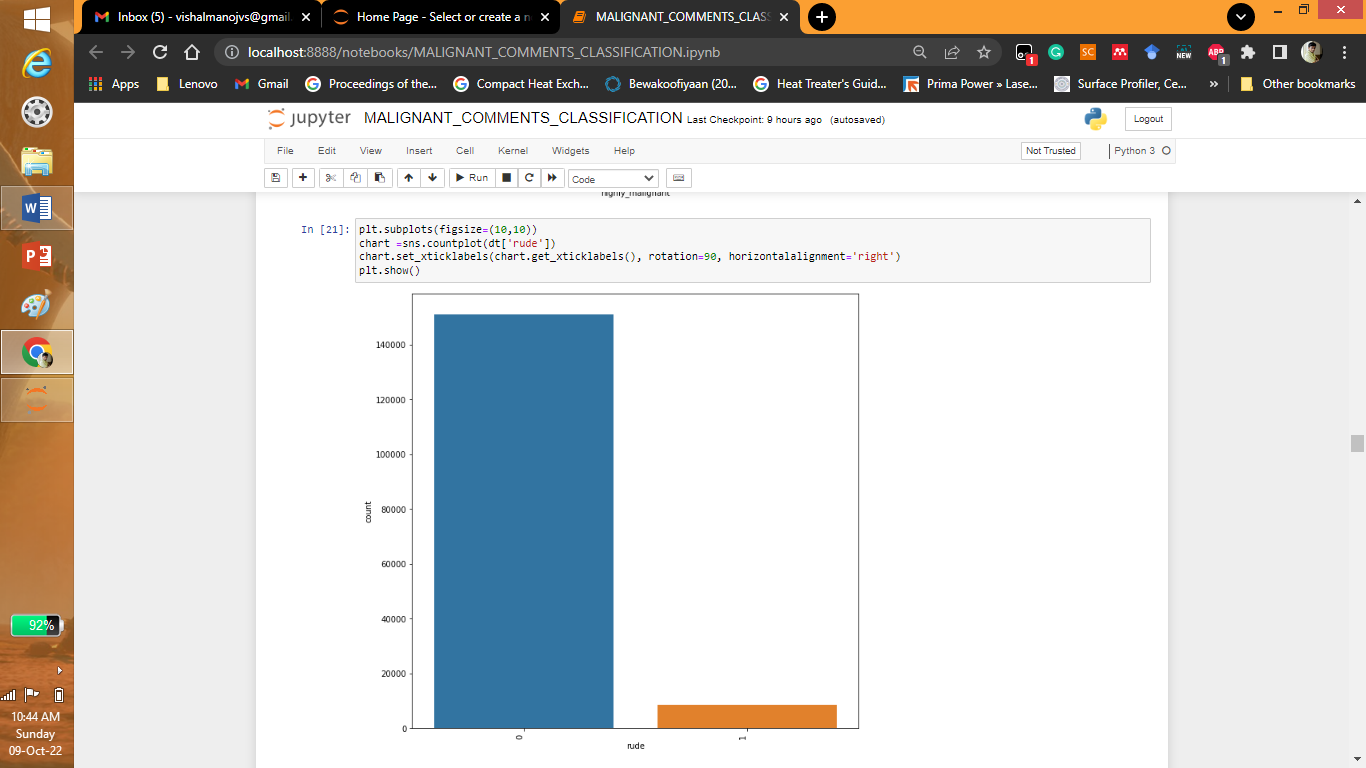
Answer: To find different images

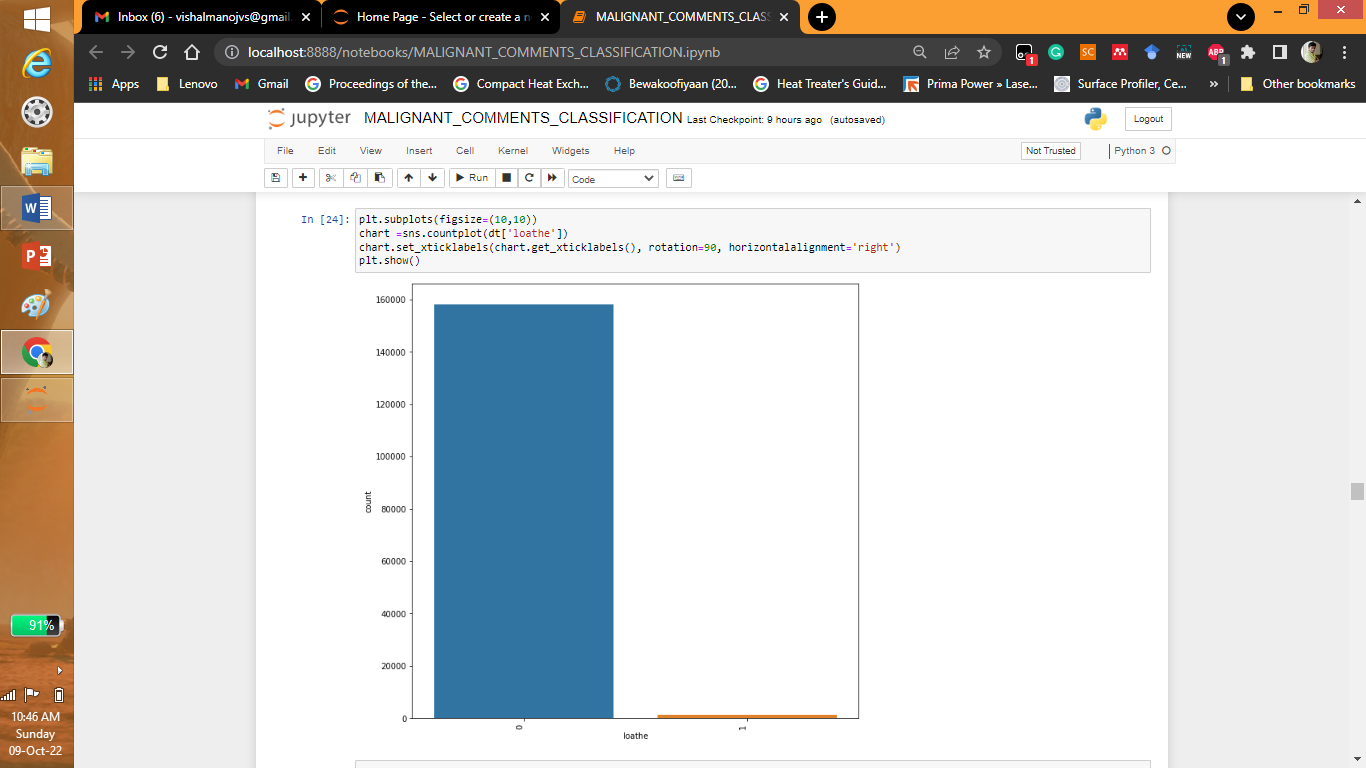
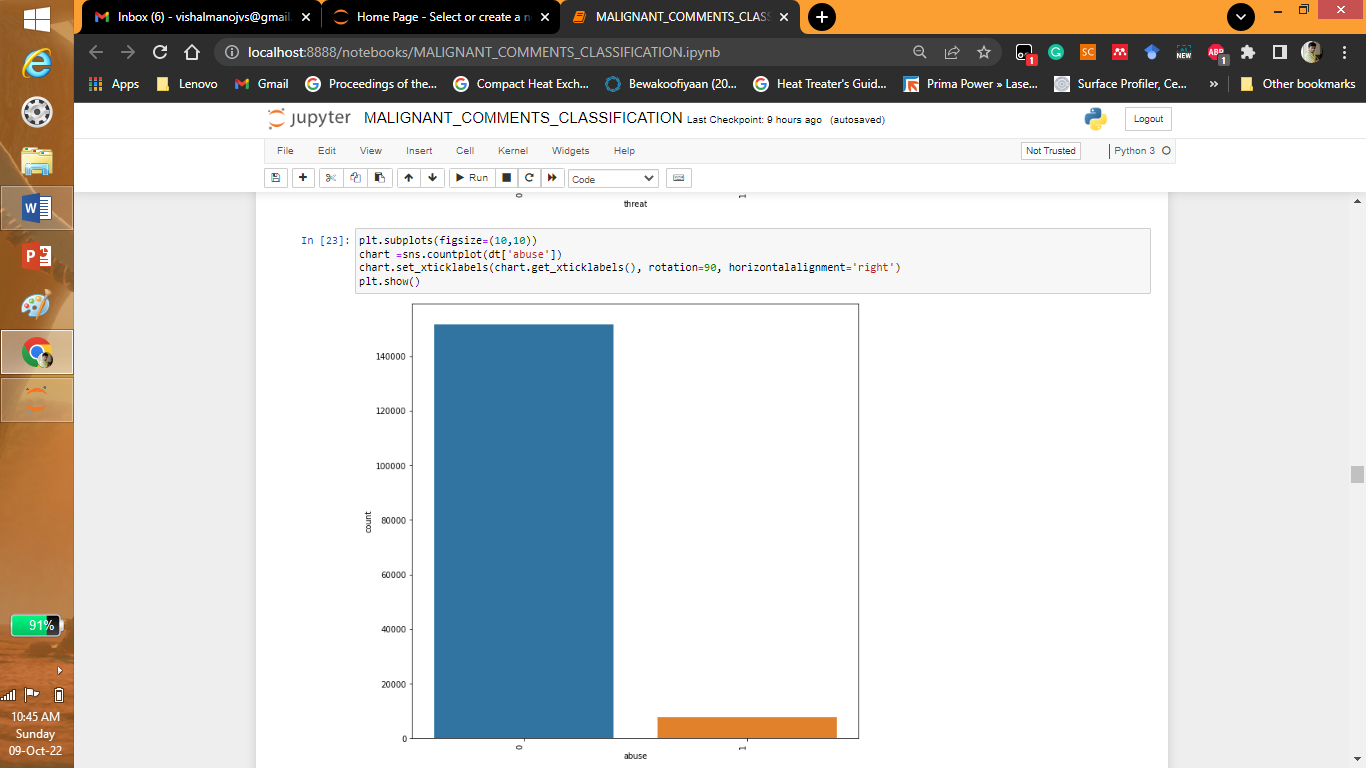


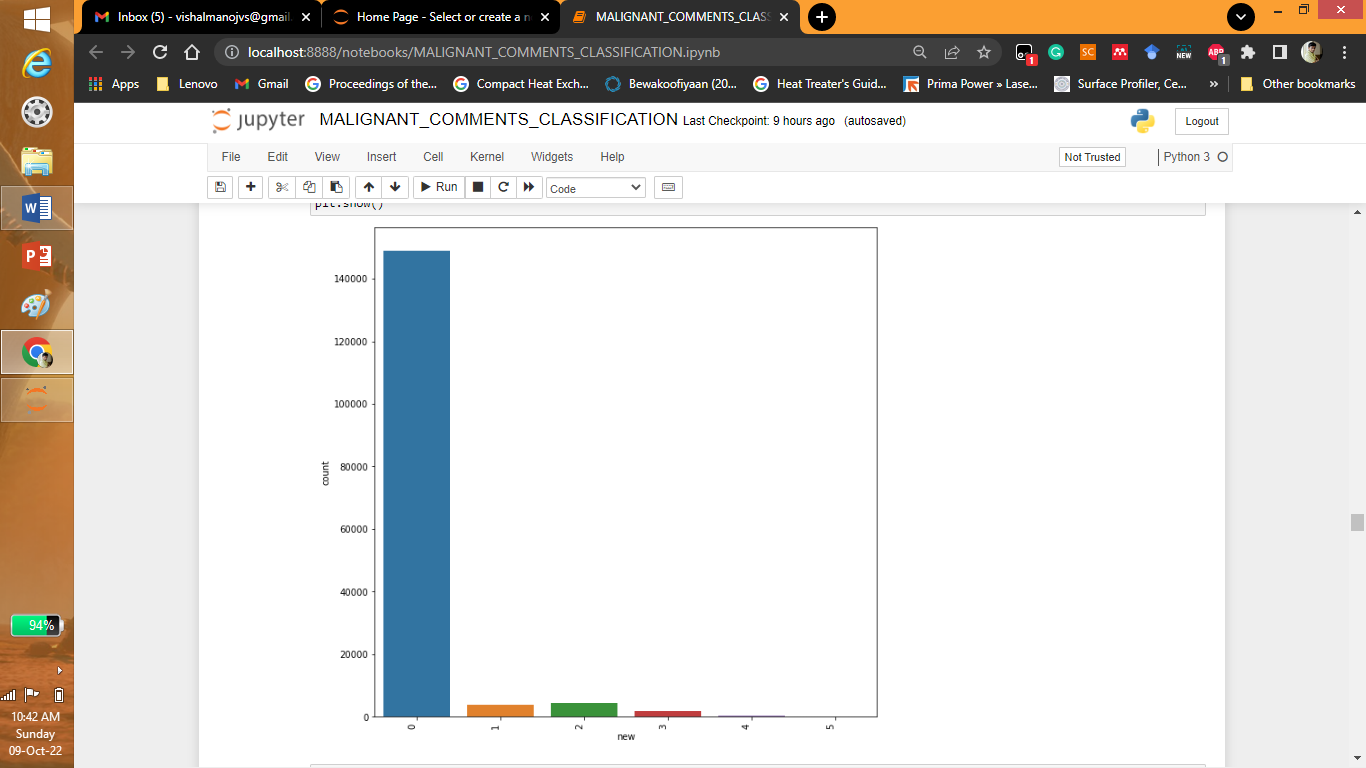




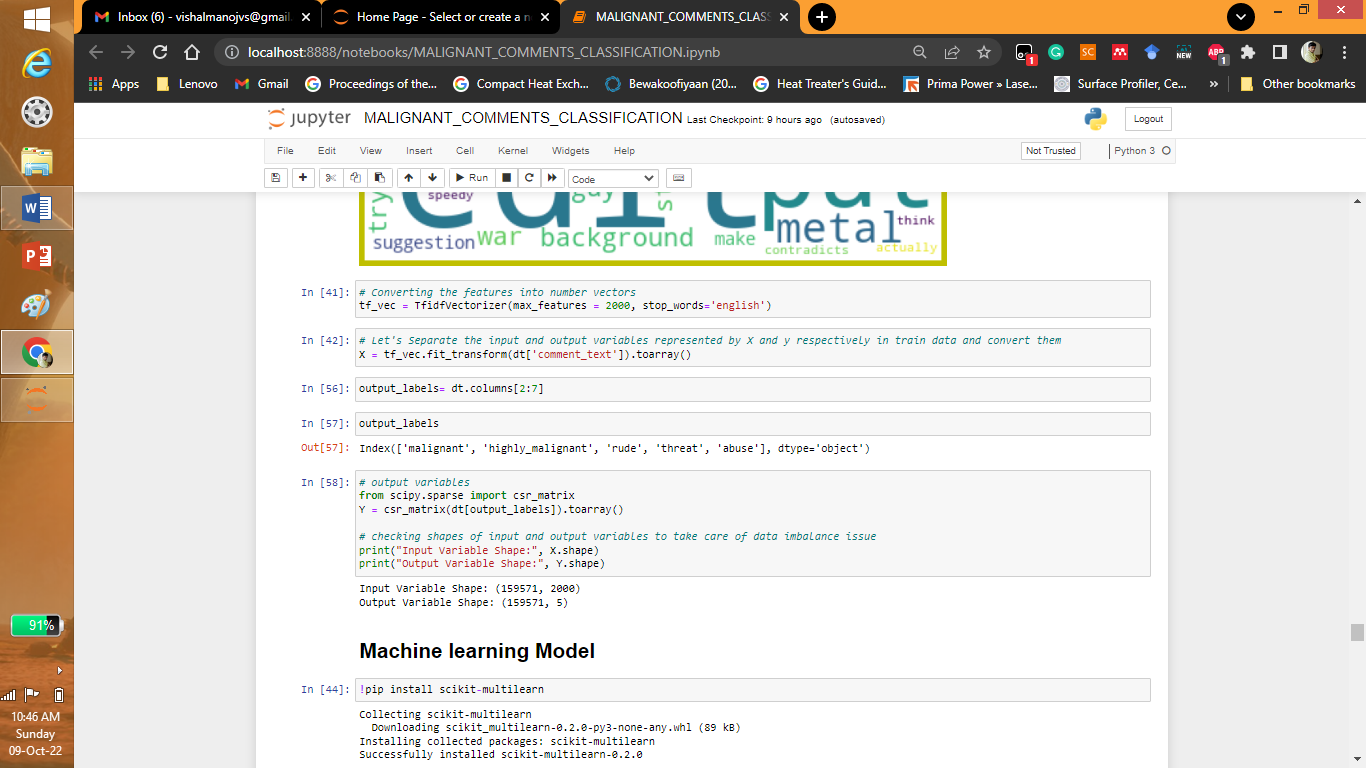








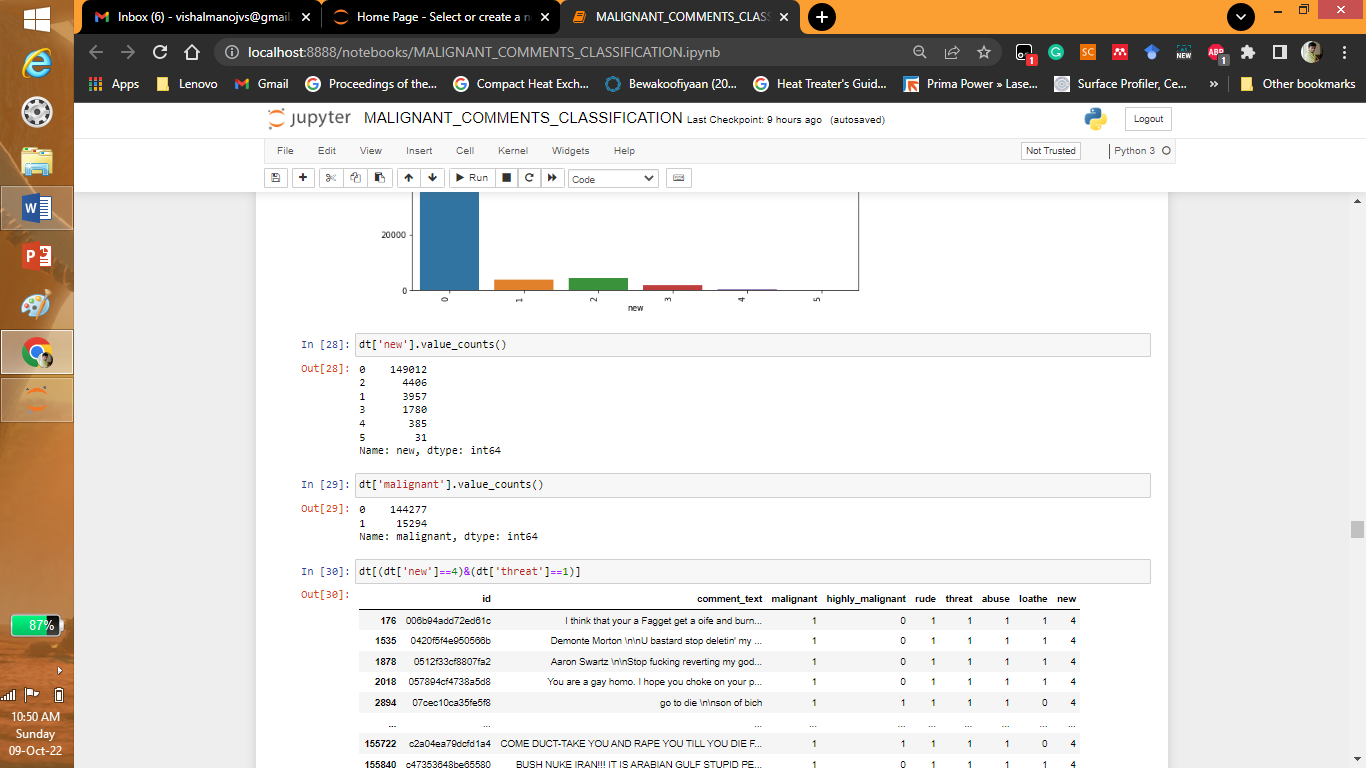




* Interpretation of the Results

Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

Answer: The neutral comments are more than other comments



The absurd, fucking… many are more common words.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Describe the key findings, inferences, observations from the whole problem.

Answer: The key finding are Linear Support Vector Classifier performs better with Accuracy Score: 91.15077857956704 % and Hamming Loss: 2.0952019242942144 % than the other classification models.

Final Model (Hyperparameter Tuning) is giving us Accuracy score of 91.26% which is slightly improved compare to earlier Accuracy score of 91.15%.

SVM classifier is fastest algorithm compare to others.

* Learning Outcomes of the Study in respect of Data Science

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

Answer: Cleaning of data and data visualization was the most challenging task no one from the organization or institute suggested proper insights for the code. But I had to look for different syntax experiment and then use for decoding.

* Limitations of this work and Scope for Future Work

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.

Answer: We have used LinearSVC having Accuracy score for the Best Model is: 91.50233957219251 and Hamming loss for the Best Model is: 2.3612967914438503. We can use many deep learning methods like ANN for better accuracy.