

NAME OF THE PROJECT

**MALIGNANT COMMENTS CLASSIFICATION**

Submitted by:

REENA

**ACKNOWLEDGMENT**

I would like to express my deepest gratitude to my SME (Subject Matter

Expert) Khushboo Garg as well as Flip Robo Technologies who gave me the

opportunity to do this project on Malignant Comments Classification,

which also helped me in doing lots of research where I came to know about

so many new things especially the Natural Language Processing and

Natural Language Toolkit parts.

Also, I have utilized a few external resources that helped me to complete

this project. I ensured that I learn from the samples and modify things

according to my project requirement. All the external resources that were

used in creating this project are listed below:

1) https://www.google.com/

2) https://www.youtube.com/

3) https://scikit-learn.org/stable/user\_guide.html

4) https://github.com/

5) https://www.kaggle.com/

6) https://medium.com/

7) https://towardsdatascience.com/

8) <https://www.analyticsvidhya.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

influencers 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 be 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**

Online platforms and social media become the place where people

share the thoughts freely without any partiality and overcoming all

the race people share their thoughts and ideas among the crowd.

Social media is a computer-based technology that facilitates the

sharing of ideas, thoughts, and information through the building of

virtual networks and communities. By design, social media is

Internet-based and gives users quick electronic communication of

content. Content includes personal information, documents, videos,

and photos. Users engage with social media via a computer, tablet,

or smartphone via web-based software or applications.

While social media is ubiquitous in America and Europe, Asian

countries like India lead the list of social media usage. More than 3.8

billion people use social media.

In this huge online platform or an online community there are some

people or some motivated mob wilfully bully others to make them

not to share their thought in rightful way. They bully others in a foul

language which among the civilized society is seen as ignominy. And

when innocent individuals are being bullied by these mob these

individuals are going silent without speaking anything. So, ideally the

motive of this disgraceful mob is achieved.

To solve this problem, we are now building a model that identifies all

the foul language and foul words, using which the online platforms

like social media principally stops these mob using the foul language

in an online community or even block them or block them from using

this foul language.

** Review of Literature**

**The purpose of the literature review is to:**

1. Identify the foul words or foul statements that are being used.

2. Stop the people from using these foul languages in online public

forum. To solve this problem, we are now building a model using our

machine language technique that identifies all the foul language and foul words, using which the online platforms like social media principally stops these mob using the foul language in an online community or even block them or block them from using this foul language.

I have used 9 different Classification algorithms and shortlisted the

best on basis of the metrics of performance and I have chosen one

algorithm and built a model in that algorithm.

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. 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.

**Motivation for the Problem Undertaken**

One of the first lessons we learn as children is that the louder you

scream and the bigger of a tantrum you throw, the more you get your

way. Part of growing up and maturing into an adult and functioning

member of society is learning how to use language and reasoning

skills to communicate our beliefs and respectfully disagree with

others, using evidence and persuasiveness to try and bring them over

to our way of thinking. Social media is reverting us back to those animalistic tantrums, schoolyard taunts and unfettered bullying that define youth, creating a dystopia where even renowned academics and dispassionate

journalists transform from Dr. Jekyll into raving Mr. Hydes, raising

the critical question of whether social media should simply enact a

blanket ban on profanity and name calling? Actually, ban should be

implemented on these profanities and taking that as a motivation I

have started this project to identify the malignant comments in social

media or in online public forums. With widespread usage of online social networks and its popularity, social networking platforms have given us incalculable opportunities than ever before, and its benefits are undeniable. Despite benefits, people may be humiliated, insulted, bullied, and harassed by

anonymous users, strangers, or peers. In this study, we have proposed a cyberbullying detection framework to generate features from online content by leveraging a pointwise mutual information technique. Based on these features, we developed a supervised machine learning solution for cyberbullying detection and multi-class categorization of its severity. Results from experiments with our proposed framework in a multi-class setting are promising both with respect to classifier accuracy and f-measure metrics. These results indicate that our proposed framework provides a feasible solution to detect cyberbullying behaviour and its severity in online social networks.

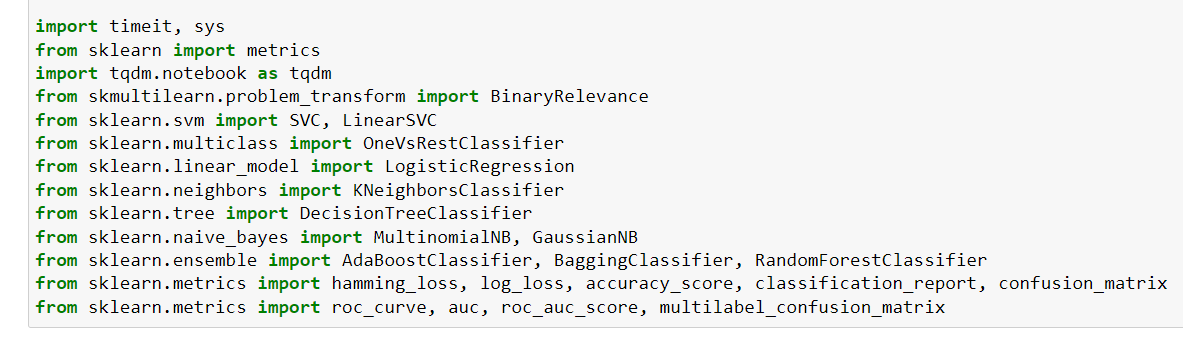
**Analytical Problem Framing**

 Mathematical/ Analytical Modelling of the Problem

The libraries/dependencies imported for this project are shown

below:





Here in this project, we have been provided with two datasets

namely train and test CSV files. I will build a machine learning model

by using NLP using train dataset. And using this model we will make

predictions for our test dataset.

I will need to build multiple classification machine learning models.

Before model building will need to perform all data pre-processing

steps involving NLP. After trying different classification models with

different hyper parameters then will select the best model out of it.

Will need to follow the complete life cycle of data science that

includes steps like -

1. Data Cleaning

2. Exploratory Data Analysis

3. Data Pre-processing

4. Model Building

5. Model Evaluation

6. Selecting the best model

Finally, we compared the results of proposed and baseline features

with other machine learning algorithms. Findings of the comparison

indicate the significance of the proposed features in cyberbullying

detection.

 **Data Sources and their format**

The data set contains the training set, which has approximately

1,59,000 samples and the test set which contains nearly 1,53,000

samples. All the data samples contain 8 fields which includes ‘Id’,

‘Comments’, ‘Malignant’, ‘Highly malignant’, ‘Rude’, ‘Threat’, ‘Abuse’

and ‘Loathe’. The label can be either 0 or 1, where 0 denotes a NO

while 1 denotes a YES. There are various comments which have

multiple labels. The first attribute is a unique ID associated with each

comment.

The data set includes:

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.

This project is more about exploration, feature engineering and

classification that can be done on this data. Since the data set is huge

and includes many categories of comments, we can do good amount

of data exploration and derive some interesting features using the

comments text column available. You need to build a model that can

differentiate between comments and its categories. **Data Pre-processing Done**

The following pre-processing pipeline is required to be performed

before building the classification model prediction:

1. Load dataset

2. Remove null values

3. Drop column id

4. Convert comment text to lower case and replace '\n' with single

space.

5. Keep only text data i.e., a-z' and remove other data from comment

text.

6. Remove stop words and punctuations

7. Apply Stemming using SnowballStemmer

8. Convert text to vectors using TfidfVectorizer

9. Load saved or serialized model

10. Predict values for multi class label

**Data Inputs- Logic- Output Relationships**

I have analysed the input output logic with word cloud and I have

word clouded the sentenced that are classified as foul language in

every category. A tag/word cloud is a novelty visual representation

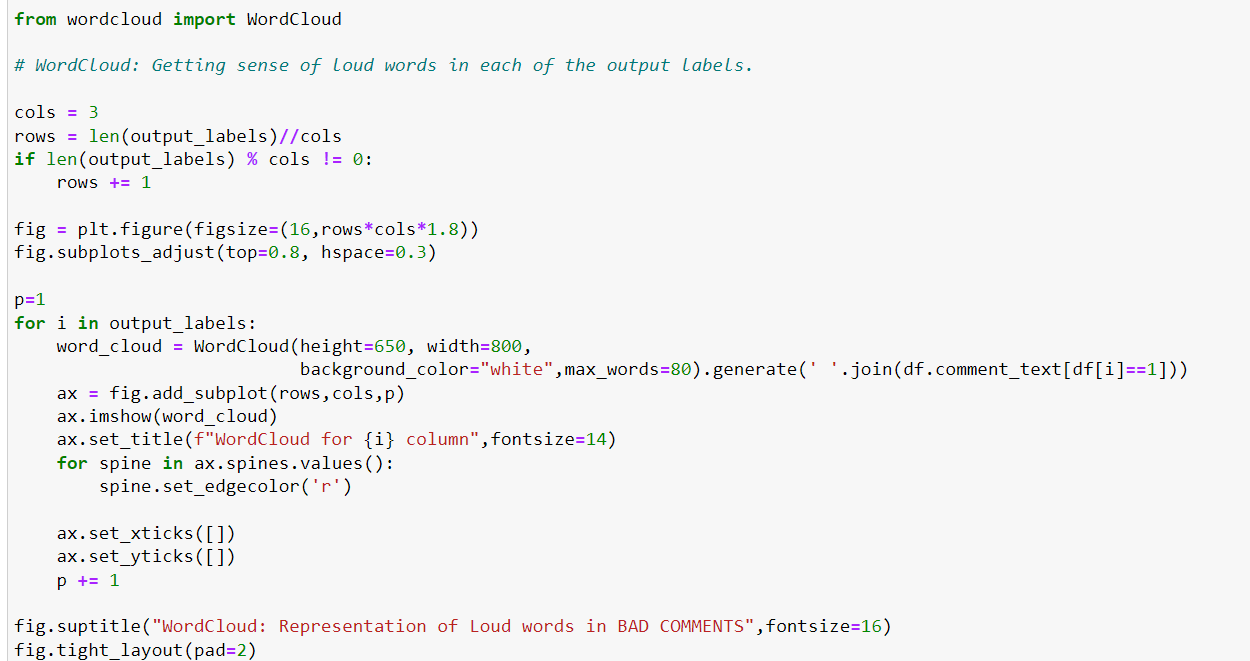
of text data, typically used to depict keyword metadata on websites,

or to visualize free form text. It’s an image composed of words used

in a particular text or subject, in which the size of each word indicates

its frequency or importance.

Code







These are the comments that belongs to different type so with which

the help of word cloud we can see if there is abuse comment which

type of words it contains to other comments as well.

** State the set of assumptions (if any) related to the**

**problem under consideration**

Cyberbullying has become a growing problem in countries around

the world. Essentially, cyberbullying doesn’t differ much from the

type of bullying that many children have unfortunately grown

accustomed to in school. The only difference is that it takes place

online.

Cyberbullying is a very serious issue affecting not just the young

victims, but also the victims' families, the bully, and those who

witness instances of cyberbullying. However, the effect of

cyberbullying can be most detrimental to the victim, of course, as

they may experience a number of emotional issues that affect their

social and academic performance as well as their overall mental

health.

**Hardware and Software Requirements and Tools Used**

Hardware technology being used.

RAM : 8.00 GB

CPU : Intel(R) Core(TM) i5-10300H CPU @ 2.50GHz

GPU : NVIDIA GeForce GTX 1650 Ti

Software technology being used.

Programming language : Python

Distribution : Anaconda Navigator

Browser based language shell : Jupyter Notebook

Libraries/Packages specifically being used.

Pandas, NumPy, matplotlib, seaborn, scikit-learn, pandas-profiling,

missingno, NLTK

**Model/s Development and Evaluation**

Identification of possible problem-solving approaches

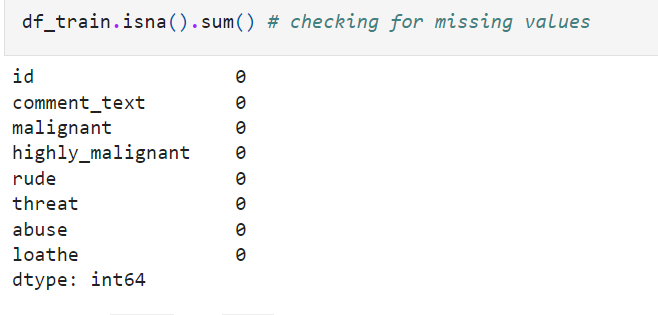
(methods)

I checked through the entire training dataset for any kind of missing

values information and all these pre-processing steps were repeated

on the testing dataset as well.

Code:





Then we went ahead and took a look at the dataset information.

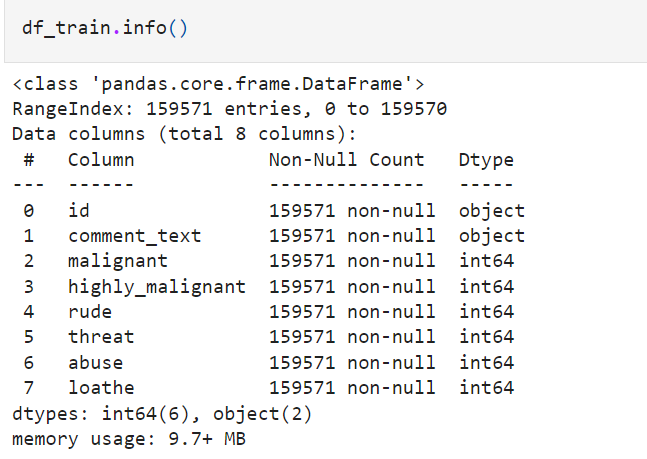
Using the info method, we are able to confirm the non-null count

details as well as the datatype information. We have a total of 8

columns out of which 2 columns have object datatype while the

remaining 6 columns are of integer datatype.

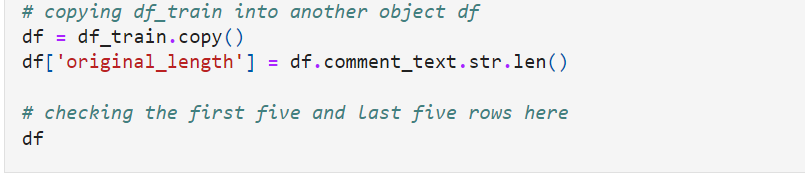
Code:



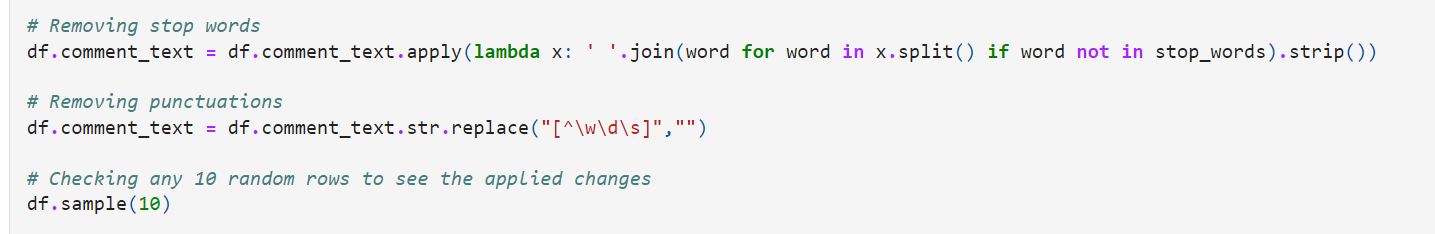
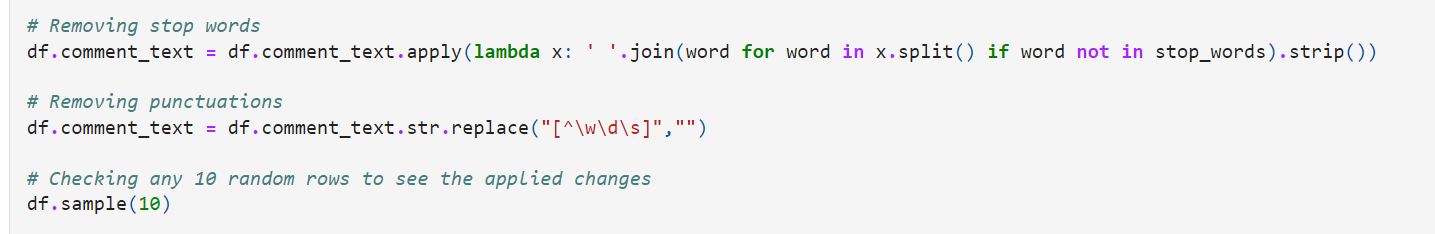
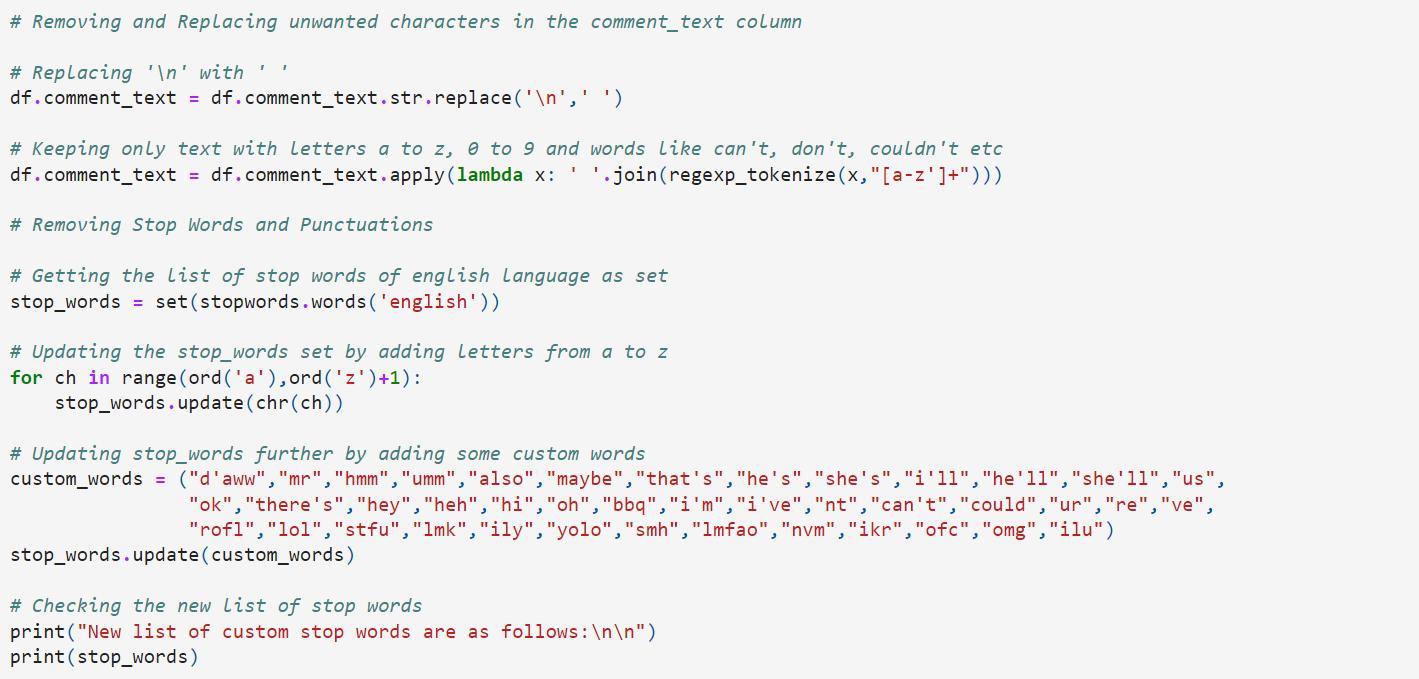
Then we went ahead and performed multiple data cleaning and data

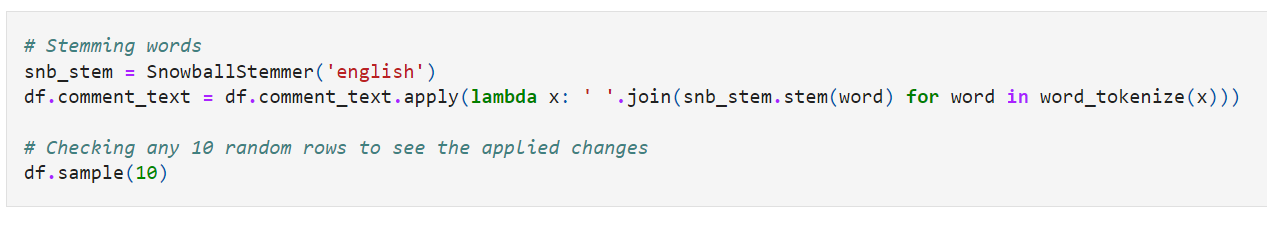
transformation steps. I have added an additional column to store the

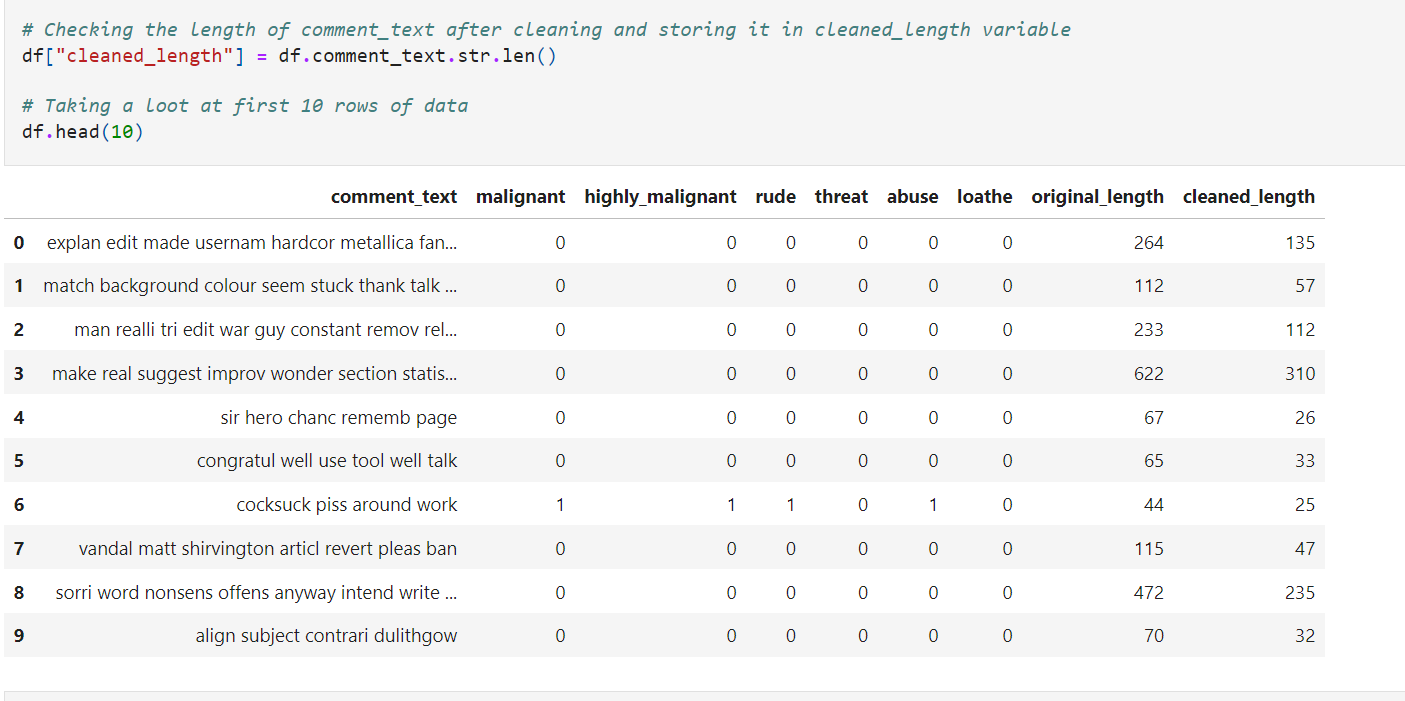
original length of our comment\_text column. 

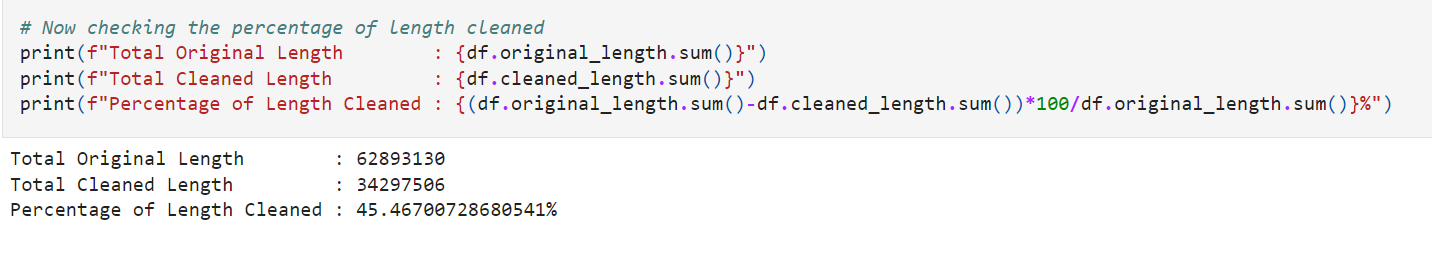












**Visualizations**

I used the pandas profiling feature to generate an initial detailed

report on my dataframe values. It gives us various information on

the rendered dataset like the correlations, missing values, duplicate

rows, variable types, memory size etc. This assists us in further

detailed visualization separating each part one by one comparing

and research for the impacts on the prediction of our target label

from all the available feature columns.

