FLIP ROBO

Malignant Comments Classification

Submitted by:

ABHISHEK MURARKA

ACKNOWLEDGMENT

I would like to express my sincere gratitude to FlipRobo Technologies for supporting me throughout the internship and giving me the opportunity to explore the depth of Data Science by

providing multiple projects like this, there are multiple people, organizations, youtubers, who guided me in this wonderful journey and few journals which helped me develop my models in this project. I would like to thank following people for the inspiration and help,

- FlipRobo Technologies (Astha Misra)
- DataTrained Team
- Jason Schvach

Research papers and youtube videos that helped me in this projects is as follows:

- https://beta.vu.nl/nl/Images/werkstuk-biere_tcm235-893877.pdf
- https://medium.com/@nupurbaghel/toxic-comment-classification-f6e075c3487a
- https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/reports/6837517.pdf

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. 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.

- ## Conceptual Background of the Domain Problem

  Classification regarding toxicity has been intensively researched in past few years, largely in the context of social media data where researchers have applied various machine learning systems to try to tackle the problem of toxicity as well as related, so there are various factors affecting the comments and can be related to the emotion, figure of speeches, and sometimes sarcasms which is related to indirect taunting is also rude and as a human it’s sometimes hard for us to distinguish between sarcasms and real appreciation.

- ## Review of Literature

  From the research paper we got to know that the we can get different approaches related to this problem, so this problem is Multilabel classification problem In multi-label classification, data can belong to more than one label simultaneously. For example, in our case a comment may be malignant, rude, threat, abuse and loathe at the same time. It may also happen that the comment is non-toxic and hence does not belong to any of the six labels, so as per research papers we have seen that the effective models for these multilabel classification problems are onevsrest claissifier, Binary Relevance Method, classifier chain Method, Adaptation Algorithm (MLKNN: This is the adapted multi label version of K-nearest neighbours. Similar to this classification algorithm is the BRkNNaClassifier and BRkNNbClassifier which are based on K-Nearest Neighbours Method. Since our problem is somewhat similar to the page categorization problem, this algorithm is expected to give acceptable results. However, the time complexity involved is large and therefore it will be preferable to train it on smaller part of the dataset.), however deep learning and Bidirectional LSTM has provided results with 96% accuracy.

- ## Motivation for the Problem Undertaken

  This project was highly motivated project as it includes the real time problem of analysing toxic behaviour and providing us opportunity to explore a bit and contribute our efforts against cyberbullying which has been proven critical as this can take a toll on anyone and affect them mentally leading to depression, mental illness, self-hatred and suicidal thoughts.

Analytical Problem Framing

- Data Sources and their formats

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.

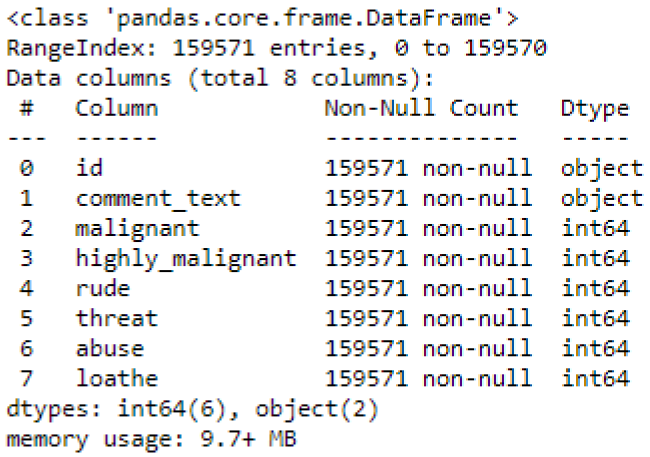
-Threat: It contains indication of the comments that are giving any threat to someone.

-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.

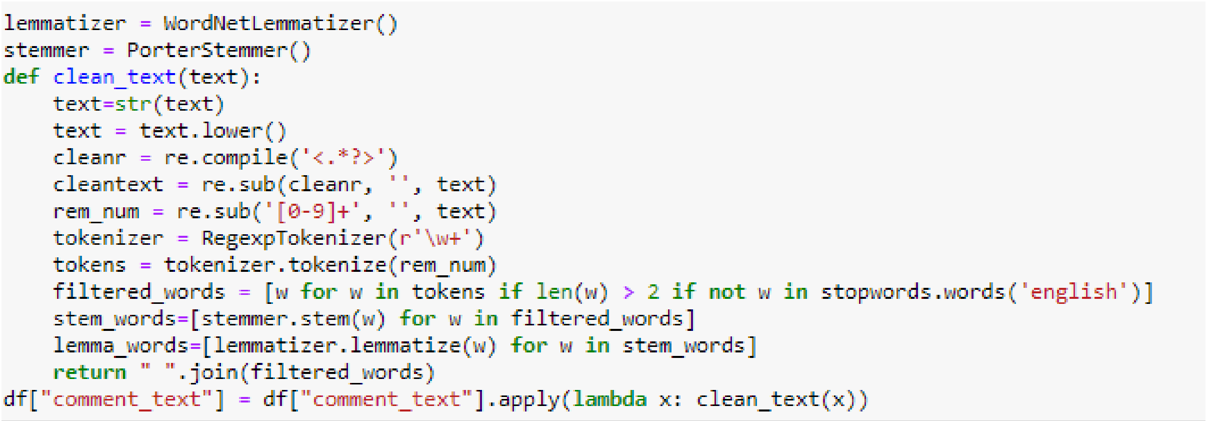
-Comment text: This column contains the comments extracted from various social media platforms.



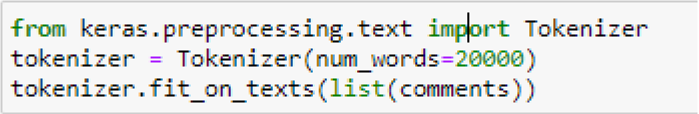
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 159571 entries, 0 to 159570
Data columns (total 8 columns):
 #   Column            Non-Null Count   Dtype 
---  ------            --------------   ----- 
 0   id                159571 non-null  object
 1   comment_text      159571 non-null  object
 2   malignant         159571 non-null  int64 
 3   highly_malignant  159571 non-null  int64 
 4   rude              159571 non-null  int64 
 5   threat            159571 non-null  int64 
 6   abuse             159571 non-null  int64 
 7   loathe            159571 non-null  int64 
dtypes: int64(6), object(2)
memory usage: 9.7+ MB
```

- ## Data Pre-processing Done

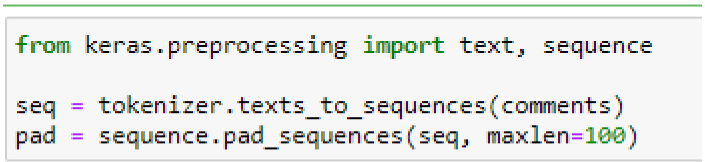
We have applied various methods for data pre processing methods in this project interestingly we use wordNet, lemmatizer and porterStemmer to clean the words and removed special characters using Regexp Tokenizer and filter the words by removing stop words and then used lemmatizes and joined and return the filtered words, Used TFIDF vectorizer to convert those text into vectors, and split the data and into test and train and trained various Machine learning algorithms, but we also explored and implemented **keras.preprocessing.text** and **keras.preprocessing** as this class allows to vectorize a text corpus, by turning each text into either a sequence of integers (each integer being the index of a token in a dictionary) or into a vector where the coefficient for each token could be binary, based on word count, based on tf-idf



```
lemmatizer = WordNetLemmatizer()
stemmer = PorterStemmer()
def clean_text(text):
    text=str(text)
    text = text.lower()
    cleanr = re.compile('<.*?>')
    cleantext = re.sub(cleanr, '', text)
    rem_num = re.sub('[0-9]+', '', text)
    tokenizer = RegexpTokenizer(r'\w+')
    tokens = tokenizer.tokenize(rem_num)
    filtered_words = [w for w in tokens if len(w) > 2 if not w in stopwords.words('english')]
    stem_words=[stemmer.stem(w) for w in filtered_words]
    lemma_words=[lemmatizer.lemmatize(w) for w in stem_words]
    return " ".join(filtered_words)
df["comment_text"] = df["comment_text"].apply(lambda x: clean_text(x))
```



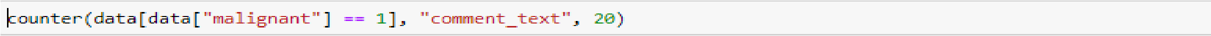
```
from keras.preprocessing.text import Tokenizer
tokenizer = Tokenizer(num_words=20000)
tokenizer.fit_on_texts(list(comments))
```



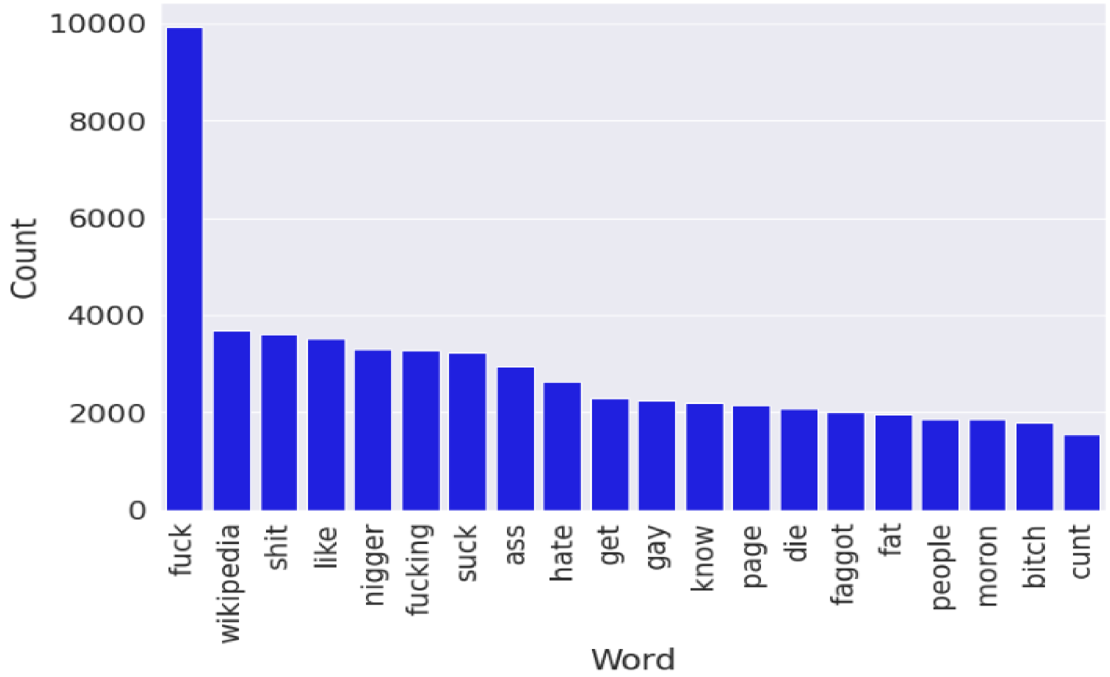
```
from keras.preprocessing import text, sequence

seq = tokenizer.texts_to_sequences(comments)
pad = sequence.pad_sequences(seq, maxlen=100)
```

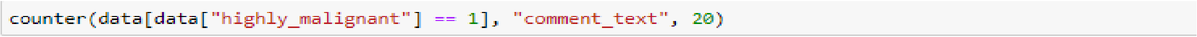
- ## Data Inputs- Logic- Output Relationships

For this data's input and output logic we will analyse words frequency for each label, so that we can get the which most 20 frequent words were used on that label categories.

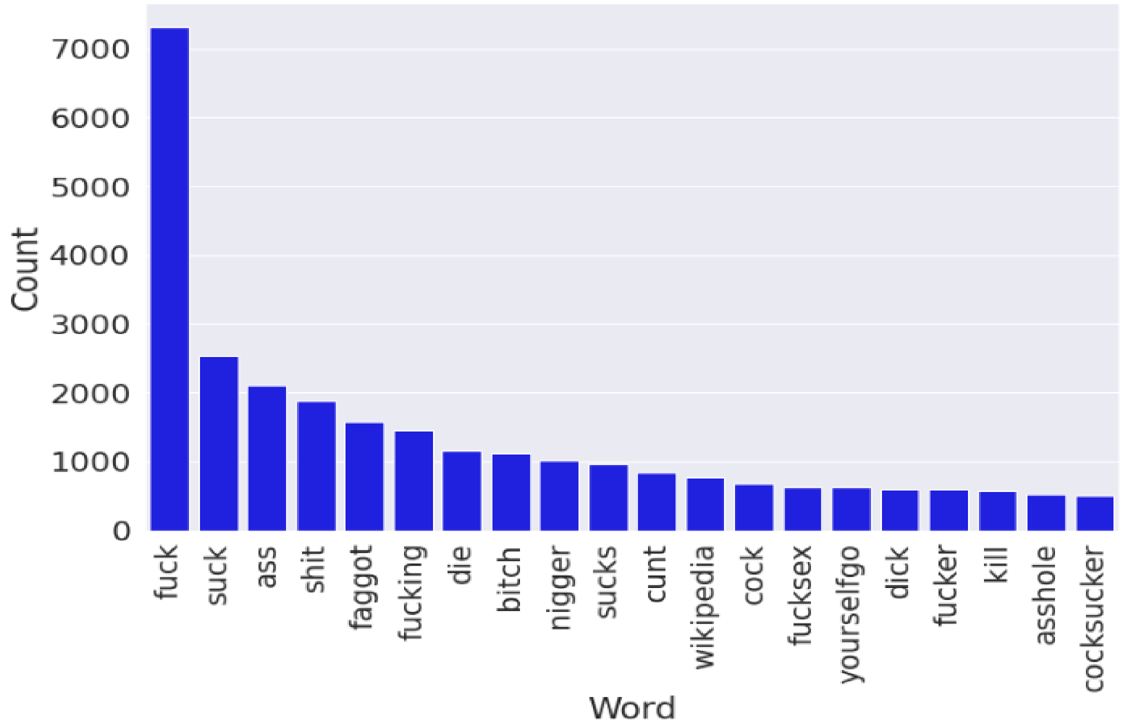


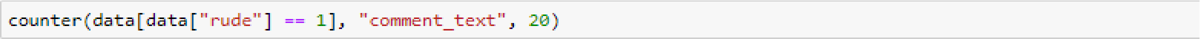
```
counter(data[data["malignant"] == 1], "comment_text", 20)
```



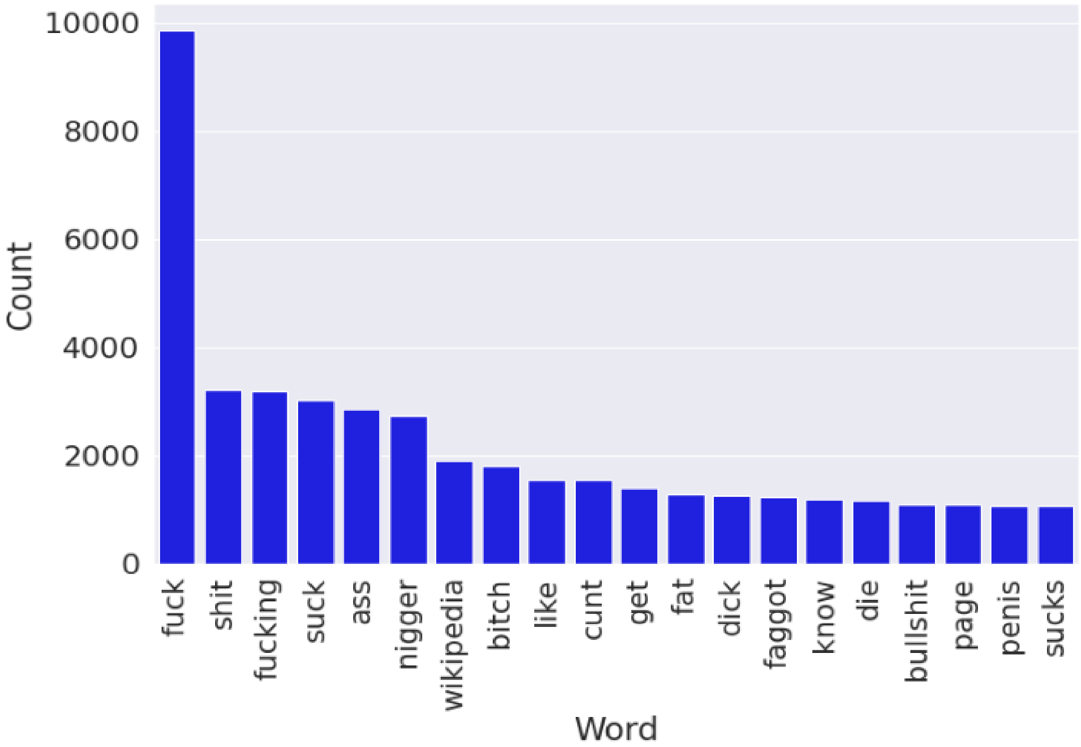


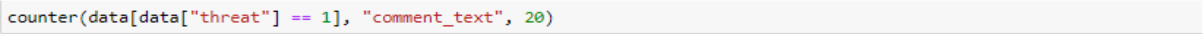
```
counter(data[data["highly_malignant"] == 1], "comment_text", 20)
```



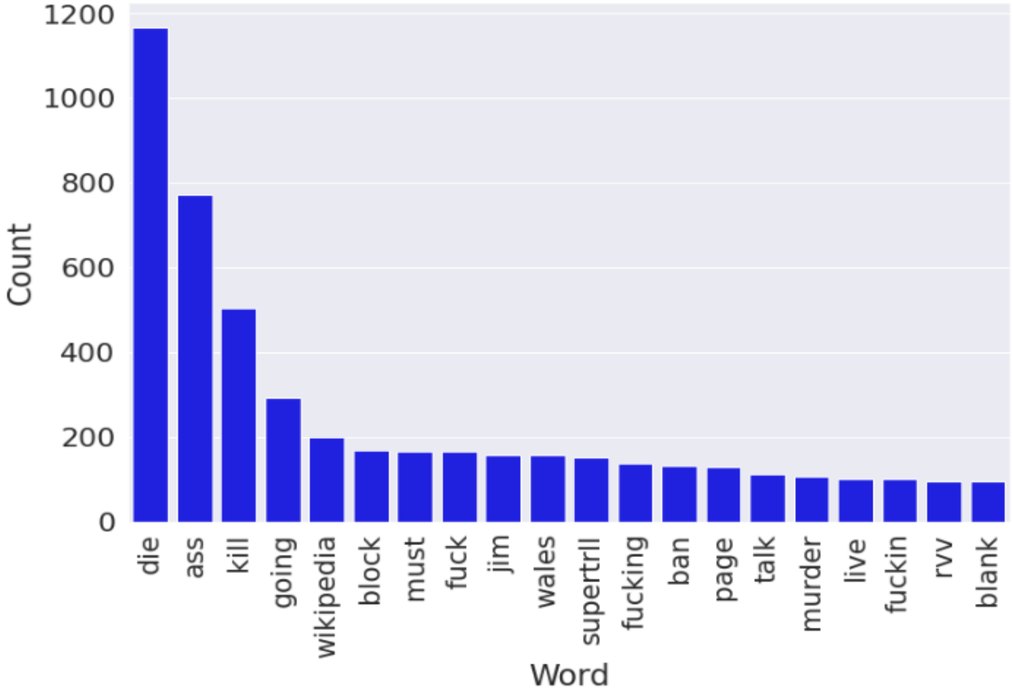


```
counter(data[data["rude"] == 1], "comment_text", 20)
```

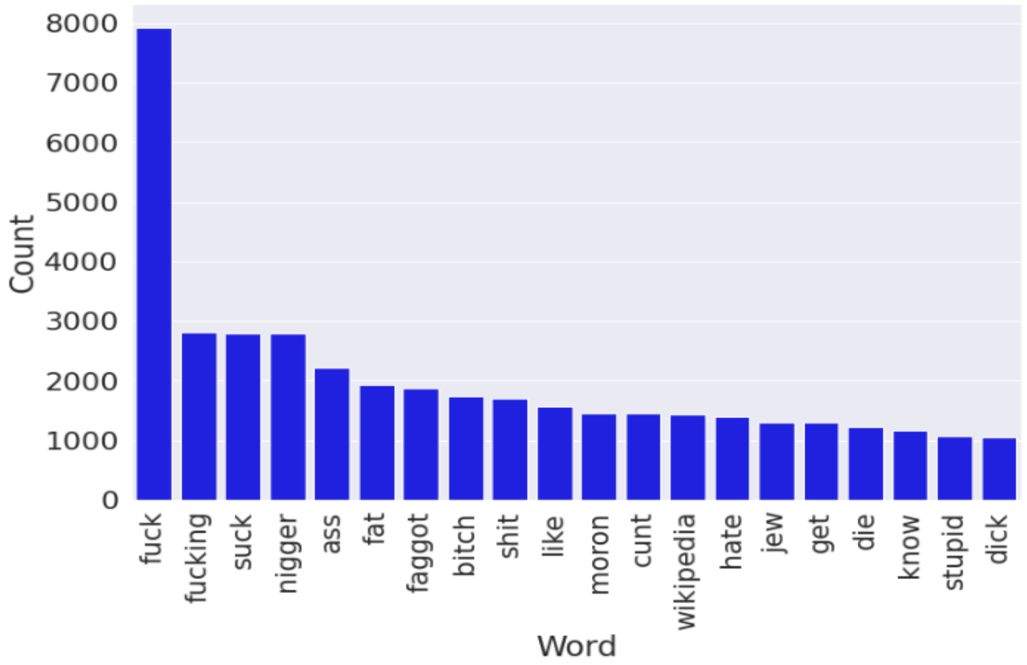




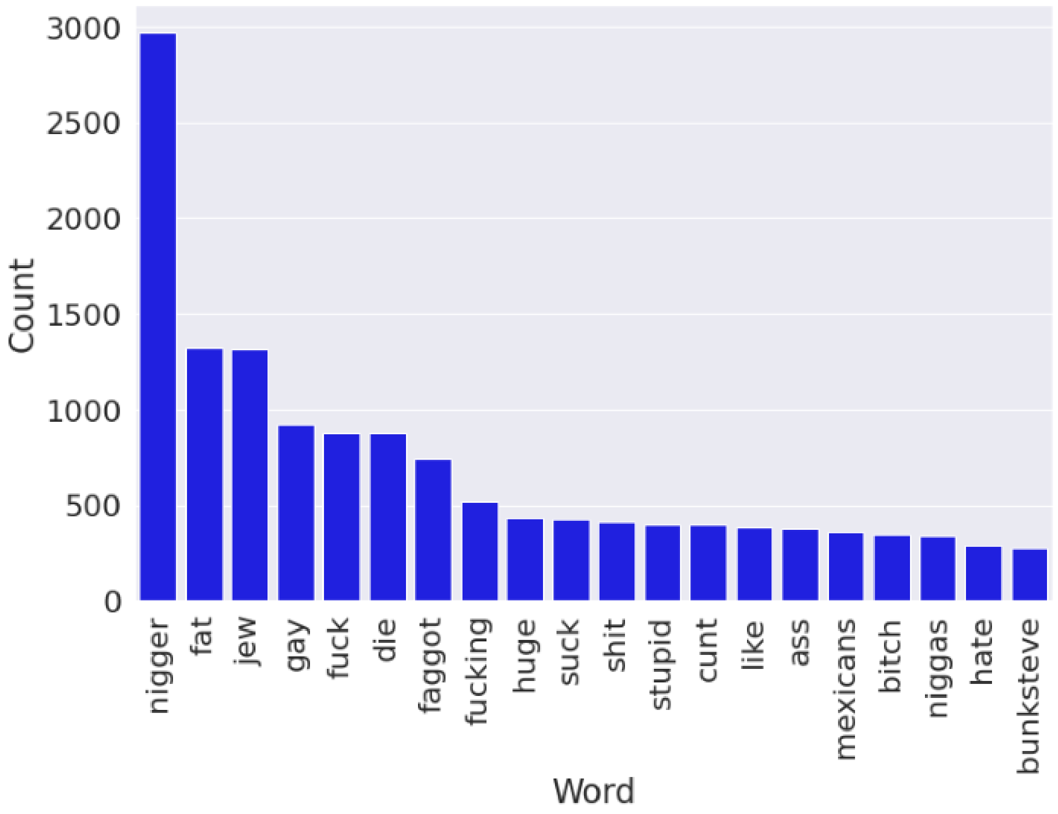
```
counter(data[data["threat"] == 1], "comment_text", 20)
```

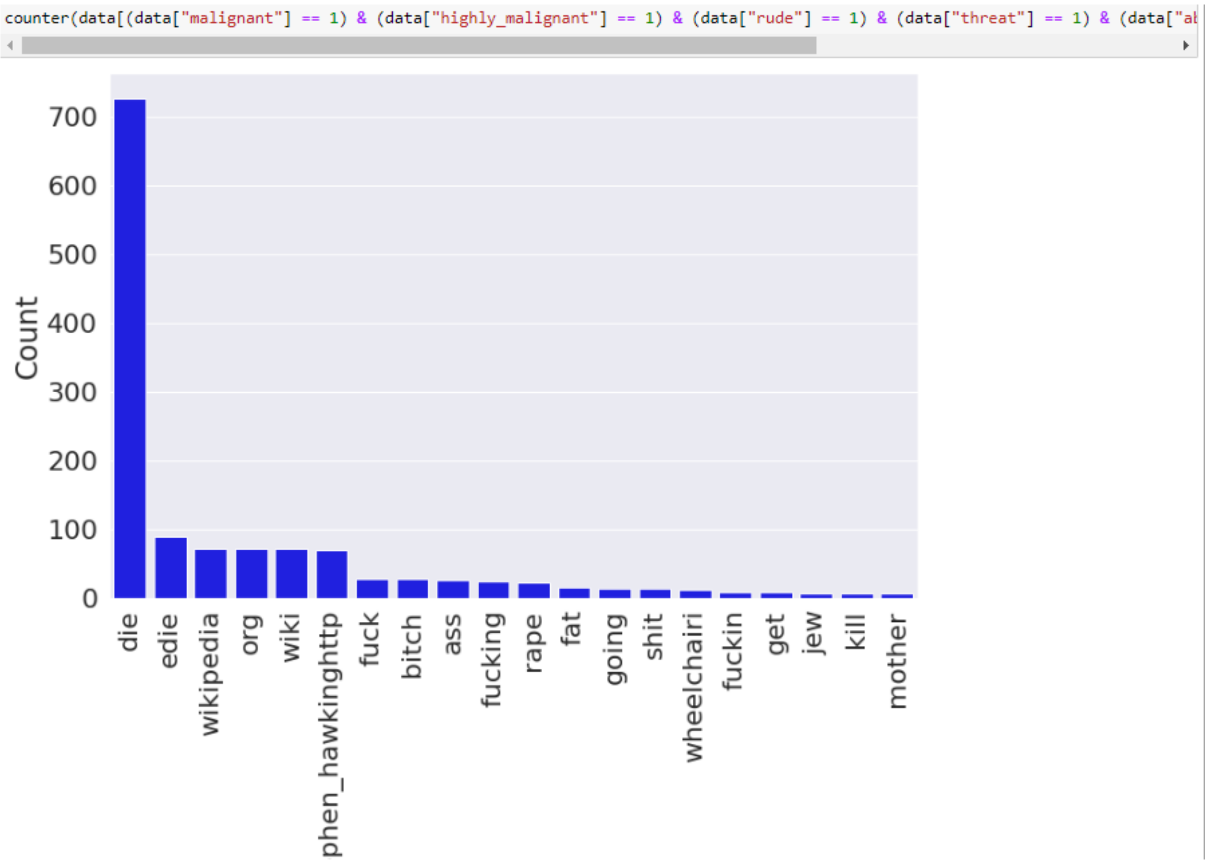


```
counter(data[data["abuse"] == 1], "comment_text", 20)
```

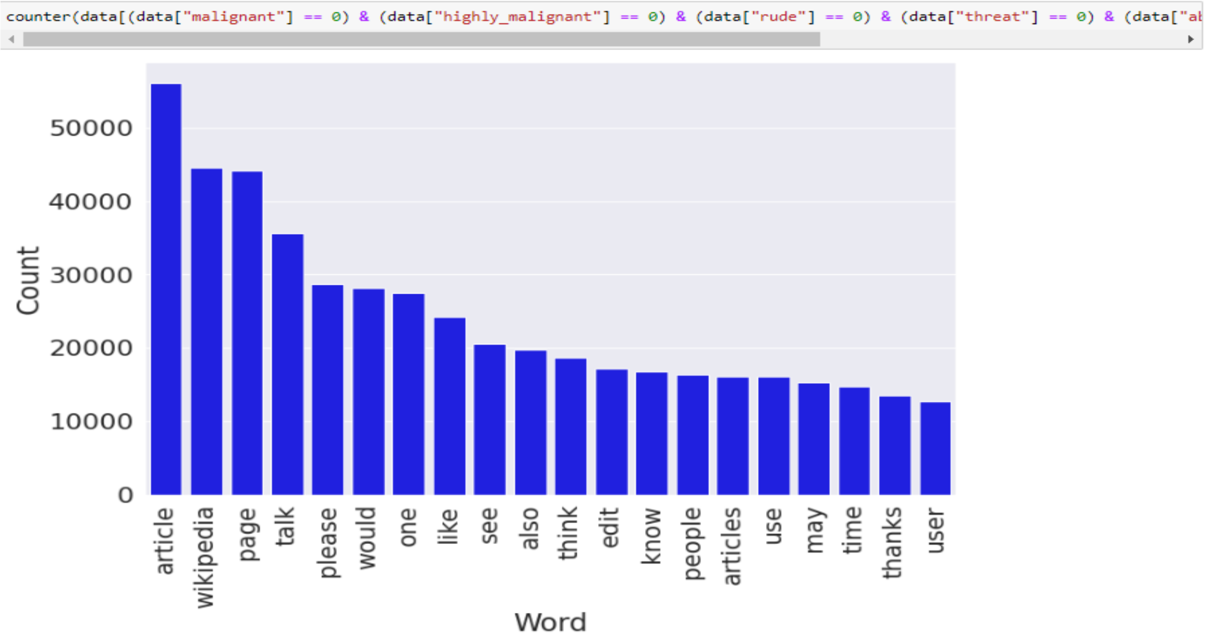


```
counter(data[data["loathe"] == 1], "comment_text", 20)
```





```
counter(data[(data["malignant"] == 1) & (data["highly_malignant"] == 1) & (data["rude"] == 1) & (data["threat"] == 1) & (data["ab
```



```
counter(data[(data["malignant"] == 0) & (data["highly_malignant"] == 0) & (data["rude"] == 0) & (data["threat"] == 0) & (data["ab
```

So from above sets of visualization we can see the most frequent words because of which that particular sentence are classified, and we can also see words which were categorized as ("malignant", "highly_malignant", "rude", "threat", "abuse”, “loathe"), and words which were completely neutral.

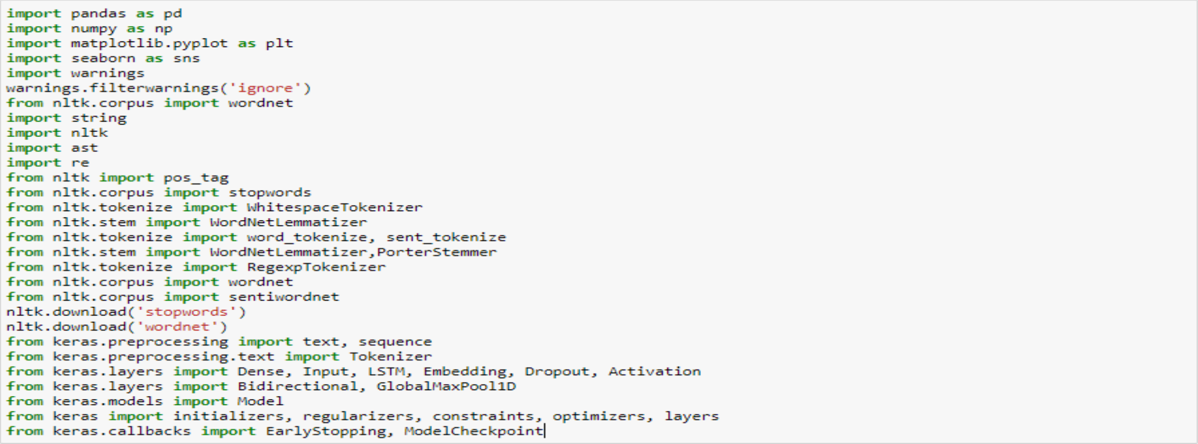
- ## State the set of assumptions (if any) related to the problem under consideration

Only assumptions which were taken related to the problem was that we dropped the id column as it had high chance of overfitting as our models could have memories the results based on id, however shuffle was used but still we couldn’t take risk.

- ## Hardware and Software Requirements and Tools Used

- Hardware: 8GB RAM, 64-bit, i7 processor, and 12 GB RAM on Googlecolab(with TPU as runtime processing)
- Software: Excel, Jupyter Notebook, python 3.6.

Library Used:

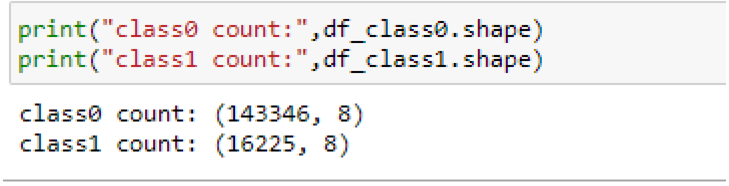


```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
from nltk.corpus import wordnet
import string
import nltk
import ast
import re
from nltk import pos_tag
from nltk.corpus import stopwords
from nltk.tokenize import WhitespaceTokenizer
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.stem import WordNetLemmatizer,PorterStemmer
from nltk.tokenize import RegexpTokenizer
from nltk.corpus import wordnet
from nltk.corpus import sentiwordnet
nltk.download('stopwords')
nltk.download('wordnet')
from keras.preprocessing import text, sequence
from keras.preprocessing.text import Tokenizer
from keras.layers import Dense, Input, LSTM, Embedding, Dropout, Activation
from keras.layers import Bidirectional, GlobalMaxPool1D
from keras.models import Model
from keras import initializers, regularizers, constraints, optimizers, layers
from keras.callbacks import EarlyStopping, ModelCheckpoint
```

Model/s Development and Evaluation

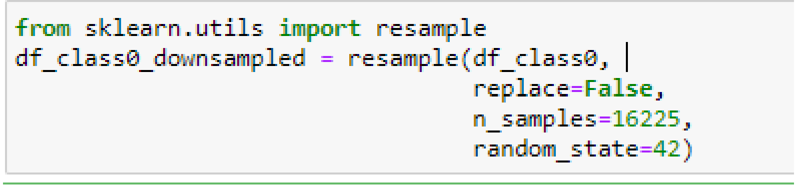
- ## Identification of possible problem-solving approaches (methods)

We also found that the dataset was highly imbalanced as class 1 had 16225, and class 0 had 143346, so we down sampled class 0 to 16225 to make the data balanced using resample.



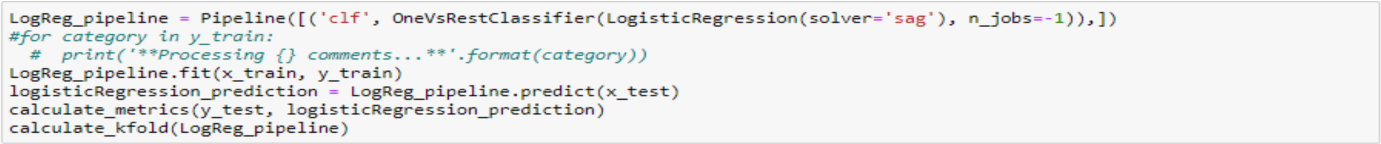
```
print("class0 count:",df_class0.shape)
print("class1 count:",df_class1.shape)
```

```
class0 count: (143346, 8)
class1 count: (16225, 8)
```

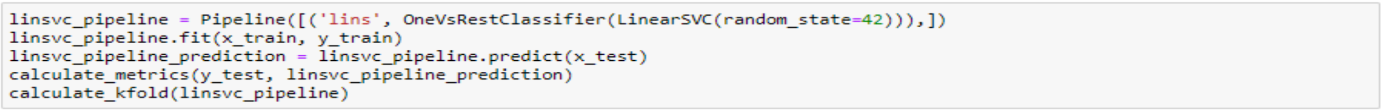


```
from sklearn.utils import resample
df_class0_downsampled = resample(df_class0, 
                                 replace=False,
                                 n_samples=16225,
                                 random_state=42)
```

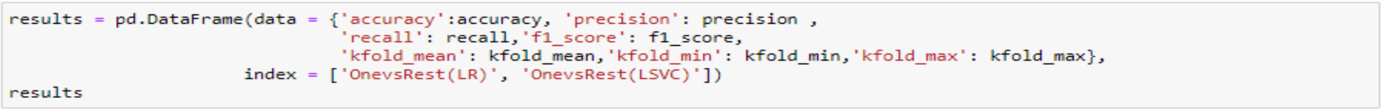
- ## Testing of Identified Approaches (Algorithms)



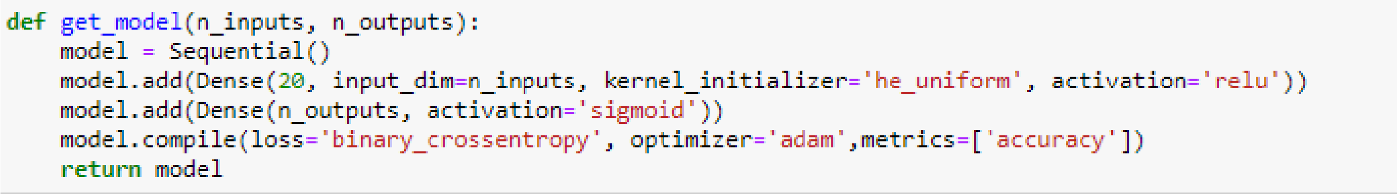
```
LogReg_pipeline = Pipeline([('clf', OneVsRestClassifier(LogisticRegression(solver='sag'), n_jobs=-1)),])
#for category in y_train:
  #  print('**Processing {} comments...**'.format(category))
LogReg_pipeline.fit(x_train, y_train)
logisticRegression_prediction = LogReg_pipeline.predict(x_test)
calculate_metrics(y_test, logisticRegression_prediction)
calculate_kfold(LogReg_pipeline)
```



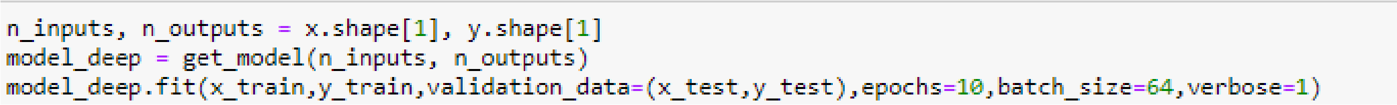
```
linsvc_pipeline = Pipeline([('lins', OneVsRestClassifier(LinearSVC(random_state=42))),])
linsvc_pipeline.fit(x_train, y_train)
linsvc_pipeline_prediction = linsvc_pipeline.predict(x_test)
calculate_metrics(y_test, linsvc_pipeline_prediction)
calculate_kfold(linsvc_pipeline)
```



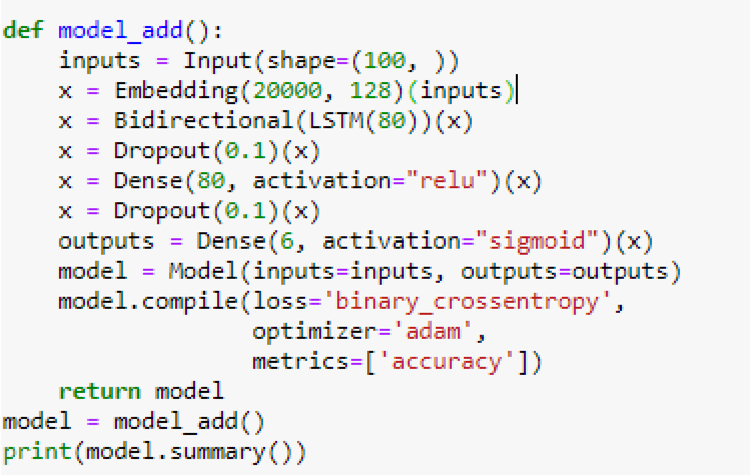
```
results = pd.DataFrame(data = {'accuracy':accuracy, 'precision': precision ,
                               'recall': recall,'f1_score': f1_score,
                               'kfold_mean': kfold_mean,'kfold_min': kfold_min,'kfold_max': kfold_max},
                       index = ['OnevsRest(LR)', 'OnevsRest(LSVC)'])
results
```



```
def get_model(n_inputs, n_outputs):
    model = Sequential()
    model.add(Dense(20, input_dim=n_inputs, kernel_initializer='he_uniform', activation='relu'))
    model.add(Dense(n_outputs, activation='sigmoid'))
    model.compile(loss='binary_crossentropy', optimizer='adam',metrics=['accuracy'])
    return model
```



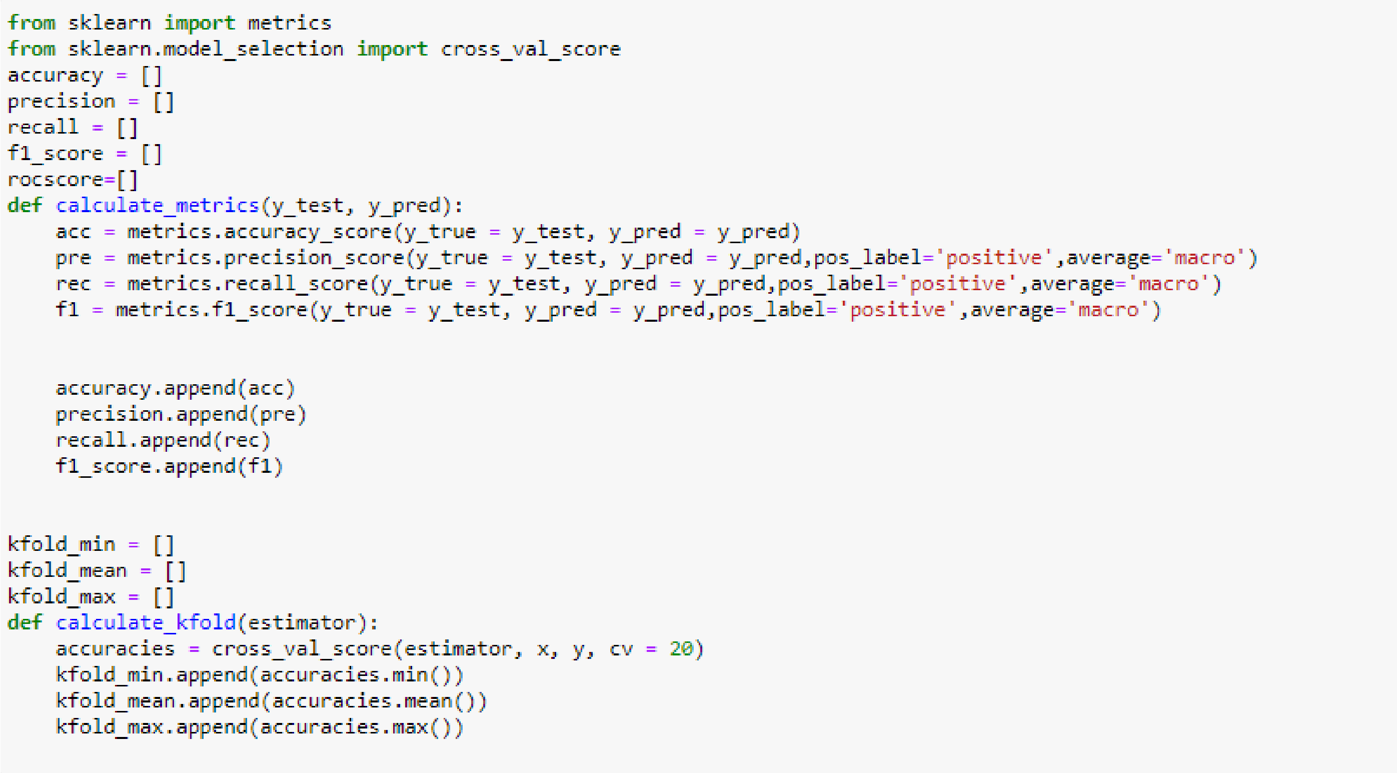
```
n_inputs, n_outputs = x.shape[1], y.shape[1]
model_deep = get_model(n_inputs, n_outputs)
model_deep.fit(x_train,y_train,validation_data=(x_test,y_test),epochs=10,batch_size=64,verbose=1)
```



```
def model_add():
    inputs = Input(shape=(100, ))
    x = Embedding(20000, 128)(inputs)
    x = Bidirectional(LSTM(80))(x)
    x = Dropout(0.1)(x)
    x = Dense(80, activation="relu")(x)
    x = Dropout(0.1)(x)
    outputs = Dense(6, activation="sigmoid")(x)
    model = Model(inputs=inputs, outputs=outputs)
    model.compile(loss='binary_crossentropy',
                  optimizer='adam',
                  metrics=['accuracy'])
    return model
model = model_add()
print(model.summary())
```

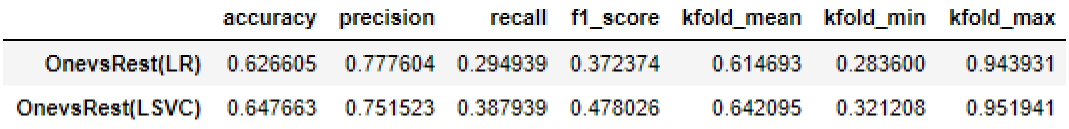
From the above we can see that the we have tried various methods and utilized pipelines in order to achieve a robust model, so we used OneVSRest classifier on logistic regression and linear SVM, and used deep learning models and bidirectional LSTM with embedding and achieve a better result.

- ## Key Metrics for success in solving problem under consideration



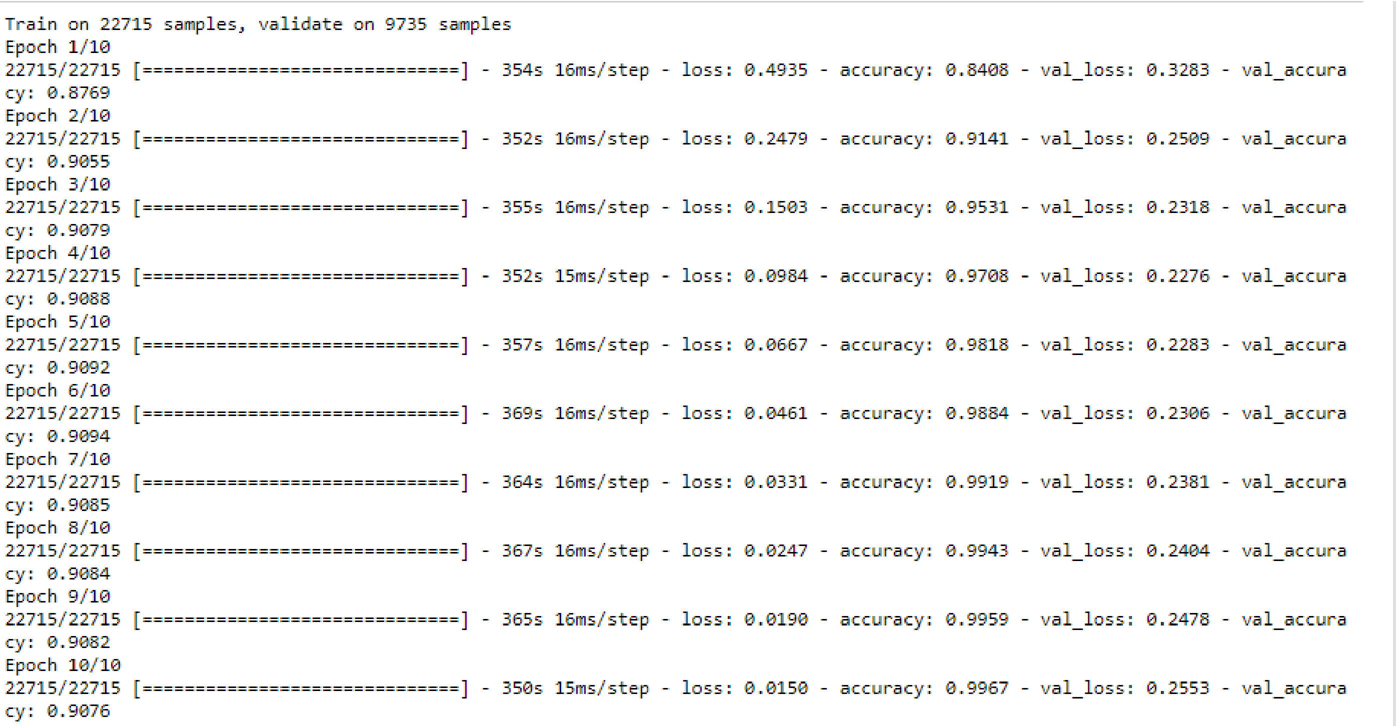
```
from sklearn import metrics
from sklearn.model_selection import cross_val_score
accuracy = []
precision = []
recall = []
f1_score = []
rocscore=[]
def calculate_metrics(y_test, y_pred):
    acc = metrics.accuracy_score(y_true = y_test, y_pred = y_pred)
    pre = metrics.precision_score(y_true = y_test, y_pred = y_pred,pos_label='positive',average='macro')
    rec = metrics.recall_score(y_true = y_test, y_pred = y_pred,pos_label='positive',average='macro')
    f1 = metrics.f1_score(y_true = y_test, y_pred = y_pred,pos_label='positive',average='macro')


    accuracy.append(acc)
    precision.append(pre)
    recall.append(rec)
    f1_score.append(f1)


kfold_min = []
kfold_mean = []
kfold_max = []
def calculate_kfold(estimator):
    accuracies = cross_val_score(estimator, x, y, cv = 20)
    kfold_min.append(accuracies.min())
    kfold_mean.append(accuracies.mean())
    kfold_max.append(accuracies.max())
```

From the above code snippet we can see the how we measured the metrices evaluation for the models which we used as OneVSRest classifier on logistic regression and linear SVC using pipelines and important metrices we needed to evaluate as it was multilabel classification model we need to use average as micro or macro but we used macro.



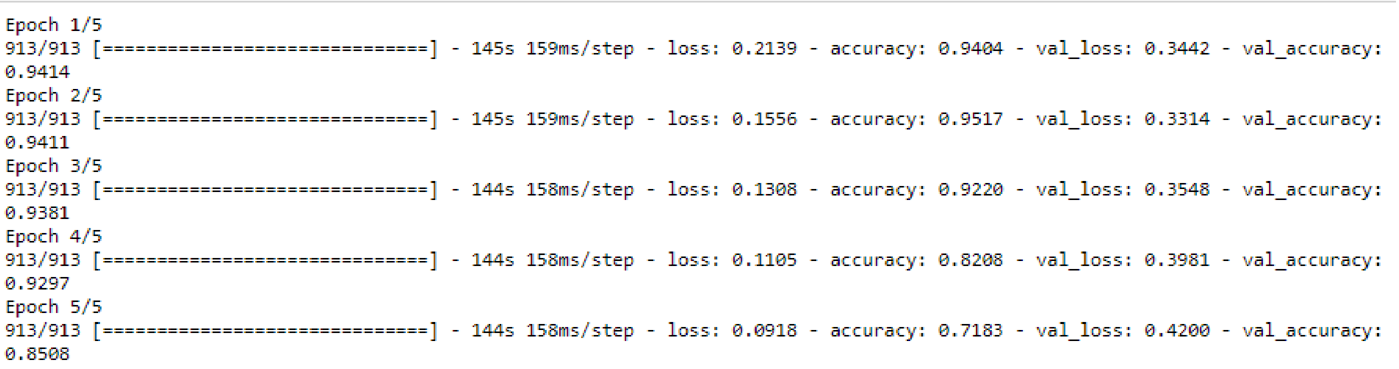
	accuracy	precision	recall	f1_score	kfold_mean	kfold_min	kfold_max
OnevsRest(LR)	0.626605	0.777604	0.294939	0.372374	0.614693	0.283600	0.943931
OnevsRest(LSVC)	0.647663	0.751523	0.387939	0.478026	0.642095	0.321208	0.951941

Deep Learning using CNN



```
Train on 22715 samples, validate on 9735 samples
Epoch 1/10
22715/22715 [==============================] - 354s 16ms/step - loss: 0.4935 - accuracy: 0.8408 - val_loss: 0.3283 - val_accura
cy: 0.8769
Epoch 2/10
22715/22715 [==============================] - 352s 16ms/step - loss: 0.2479 - accuracy: 0.9141 - val_loss: 0.2509 - val_accura
cy: 0.9055
Epoch 3/10
22715/22715 [==============================] - 355s 16ms/step - loss: 0.1503 - accuracy: 0.9531 - val_loss: 0.2318 - val_accura
cy: 0.9079
Epoch 4/10
22715/22715 [==============================] - 352s 15ms/step - loss: 0.0984 - accuracy: 0.9708 - val_loss: 0.2276 - val_accura
cy: 0.9088
Epoch 5/10
22715/22715 [==============================] - 357s 16ms/step - loss: 0.0667 - accuracy: 0.9818 - val_loss: 0.2283 - val_accura
cy: 0.9092
Epoch 6/10
22715/22715 [==============================] - 369s 16ms/step - loss: 0.0461 - accuracy: 0.9884 - val_loss: 0.2306 - val_accura
cy: 0.9094
Epoch 7/10
22715/22715 [==============================] - 364s 16ms/step - loss: 0.0331 - accuracy: 0.9919 - val_loss: 0.2381 - val_accura
cy: 0.9085
Epoch 8/10
22715/22715 [==============================] - 367s 16ms/step - loss: 0.0247 - accuracy: 0.9943 - val_loss: 0.2404 - val_accura
cy: 0.9084
Epoch 9/10
22715/22715 [==============================] - 365s 16ms/step - loss: 0.0190 - accuracy: 0.9959 - val_loss: 0.2478 - val_accura
cy: 0.9082
Epoch 10/10
22715/22715 [==============================] - 350s 15ms/step - loss: 0.0150 - accuracy: 0.9967 - val_loss: 0.2553 - val_accura
cy: 0.9076
```

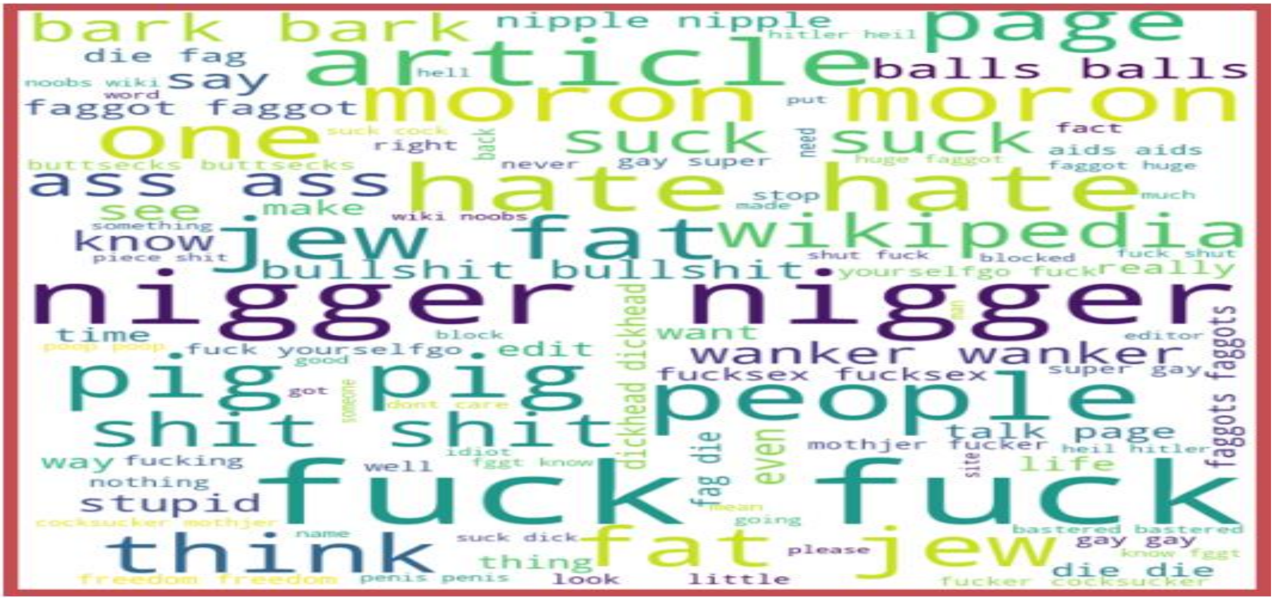
Bidirectional LSTM



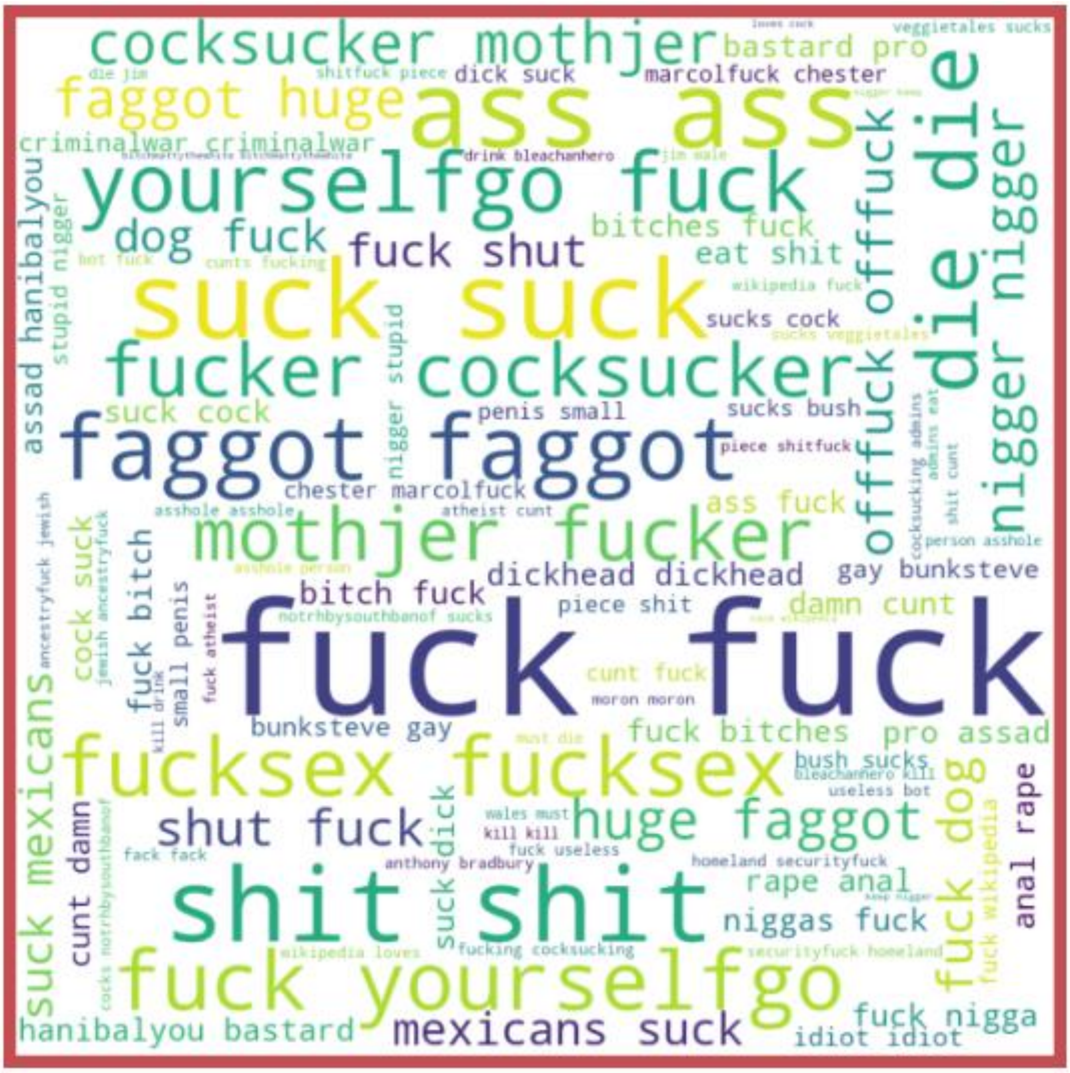
```
Epoch 1/5
913/913 [==============================] - 145s 159ms/step - loss: 0.2139 - accuracy: 0.9404 - val_loss: 0.3442 - val_accuracy:
0.9414
Epoch 2/5
913/913 [==============================] - 145s 159ms/step - loss: 0.1556 - accuracy: 0.9517 - val_loss: 0.3314 - val_accuracy:
0.9411
Epoch 3/5
913/913 [==============================] - 144s 158ms/step - loss: 0.1308 - accuracy: 0.9220 - val_loss: 0.3548 - val_accuracy:
0.9381
Epoch 4/5
913/913 [==============================] - 144s 158ms/step - loss: 0.1105 - accuracy: 0.8208 - val_loss: 0.3981 - val_accuracy:
0.9297
Epoch 5/5
913/913 [==============================] - 144s 158ms/step - loss: 0.0918 - accuracy: 0.7183 - val_loss: 0.4200 - val_accuracy:
0.8508
```

- Visualizations

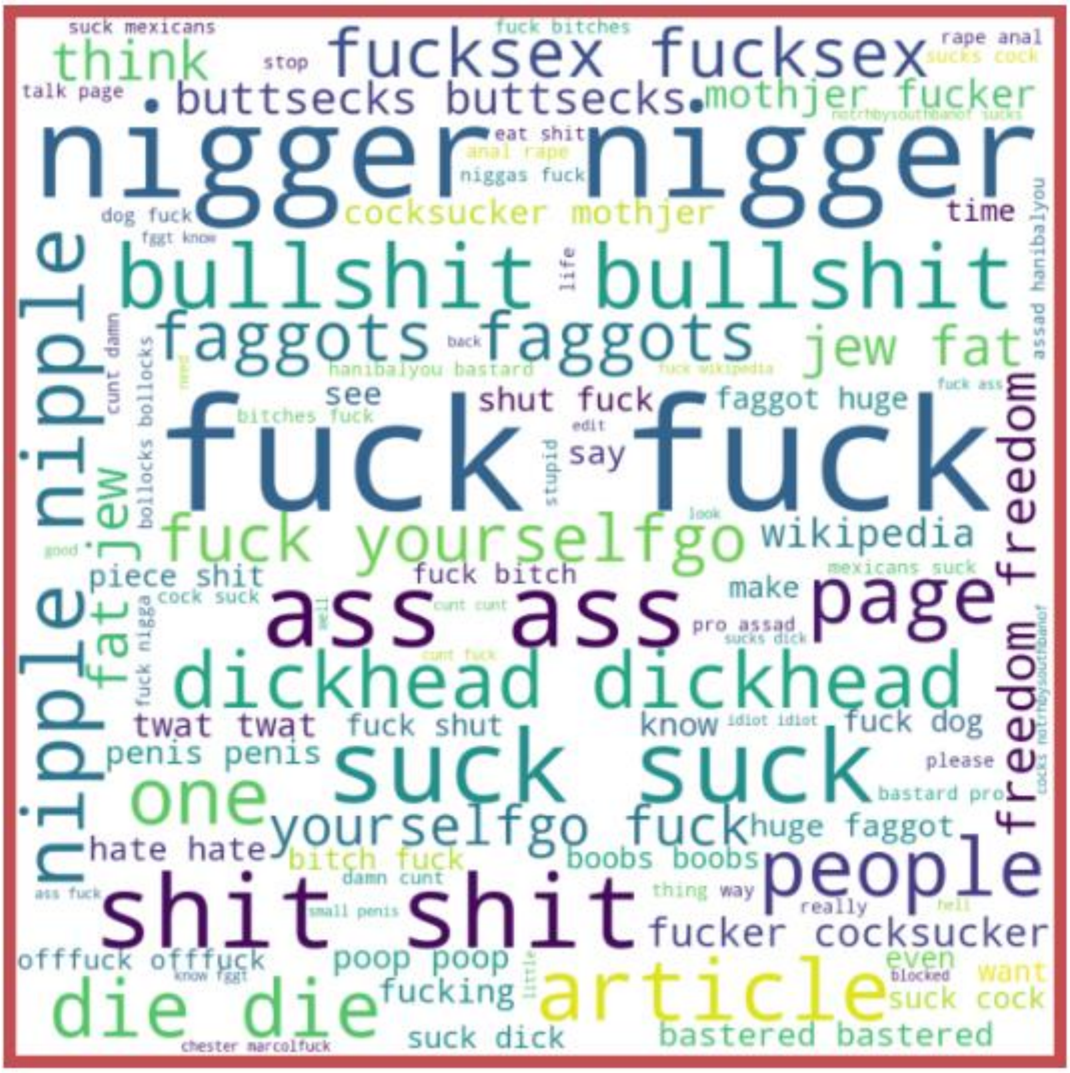
Malignanat



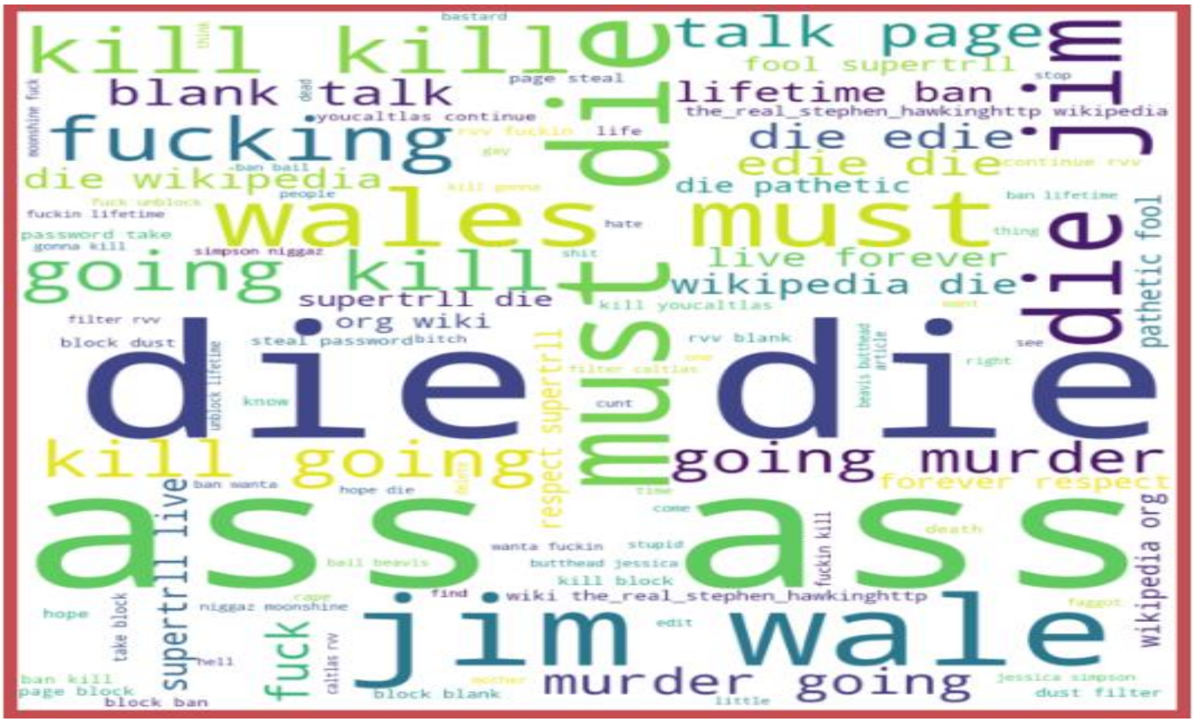
highly_malignant



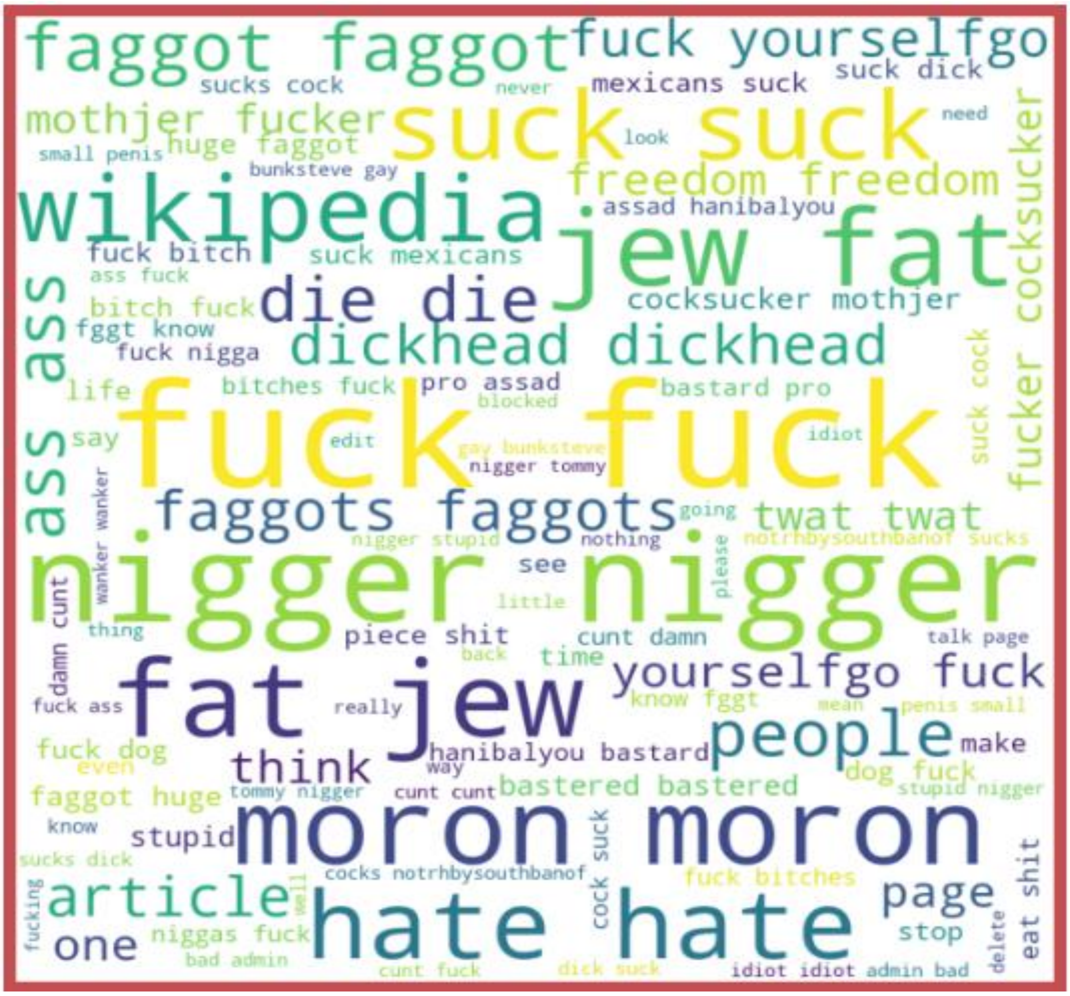
Rude


suck mexicans
fuck bitches
rape anal
think
stop
fucksex fucksex
sucks cock
talk page
buttsecks buttsecks
mothjer fucker
notrhbysouthbanof sucks
eat shit
anal rape
nigger nigger
niggas fuck
dog fuck
cocksucker mothjer
time
fggt know
assad hanibalyou
life
bullshit bullshit
nipple nipple
faggots faggots
jew fat
back
cunt damn
bollocks bollocks
hanibalyou bastard
fuck wikipedia
fuck ass
see
shut fuck
faggot huge
bitches fuck
edit
fuck fuck
freedom freedom
say
stupid
look
fat jew
good
fuck yourselfgo
wikipedia
mexicans suck
piece shit
fuck bitch
cock suck
make
ass ass
page
cunt cunt
pro assad
sucks dick
fuck nigga
cunt fuck
cocks notrhbysouthbanof
dickhead dickhead
twat twat
fuck shut
know
idiot idiot
fuck dog
penis penis
please
suck suck
one
bastard pro
yourselfgo fuck
huge faggot
hate hate
bitch fuck
boobs boobs
people
ass fuck
damn cunt
thing
way
small penis
really
hell
shit shit
fucker cocksucker
offfuck offfuck
poop poop
even
blocked
know fggt
want
little
fucking
article
suck cock
die die
suck dick
bastered bastered
chester marcolfuck

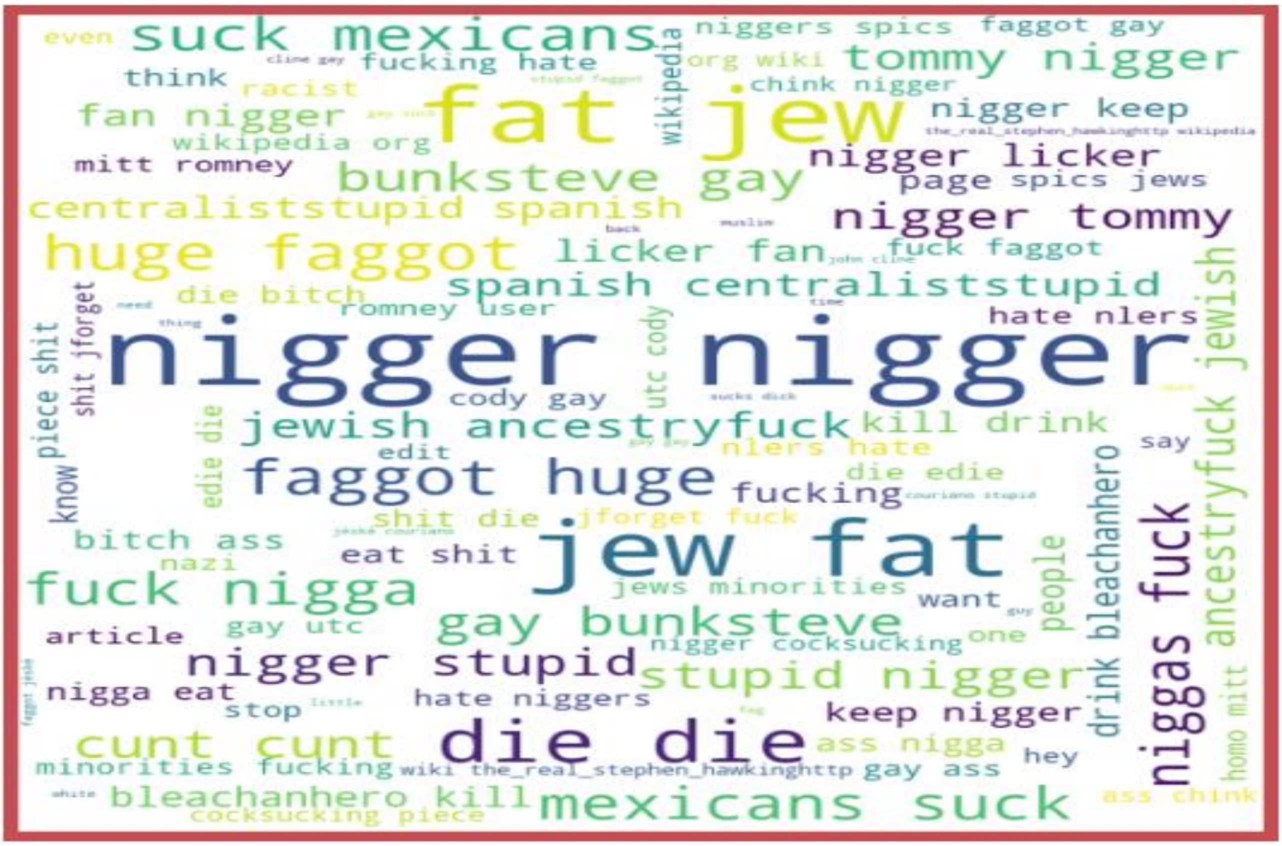
Threat


bastard
think
kill kill
die die
talk page
jim jim
fool supertrll
page steal
stop
moonshine fuck
dead
blank talk
lifetime ban
youcaltlas continue
the_real_stephen_hawkinghttp wikipedia
fucking
rvv fuckin
life
die edie
gev
edie die
continue rvv
ban ball
die wikipedia
kill gonna
die pathetic
people
fuck unblock
ban lifetime
fuckin lifetime
wales must
hate
thing
password take
pathetic fool
gonna kill
simpson niggaz
shit
live forever
going kill
must die
die die
wikipedia die
supertrll die
kill youcaltlas
want
filter rvv
org wiki
block dust
steal passwordbitch
rvv blank
see
unblock lifetime
supertrll
beavis butthead
article
filter caltlas
right
know
cunt
die die
kill going
going murder
hope die
respect supertrll
ban wanta
forever respect
time
come
supertrll live
fuckin kill
ass ass
wanta fuckin
stupid
death
wikipedia org
ball beavis
butthead jessica
rape
find
wiki the_real_stephen_hawkinghttp
niggaz moonshine
kill block
faggot
hope
take block
edit
jim wale
caltlas rvv
fuck
hell
mother
murder going
ban kill
jessica simpson
page block
block blank
little
dust filter
block ban

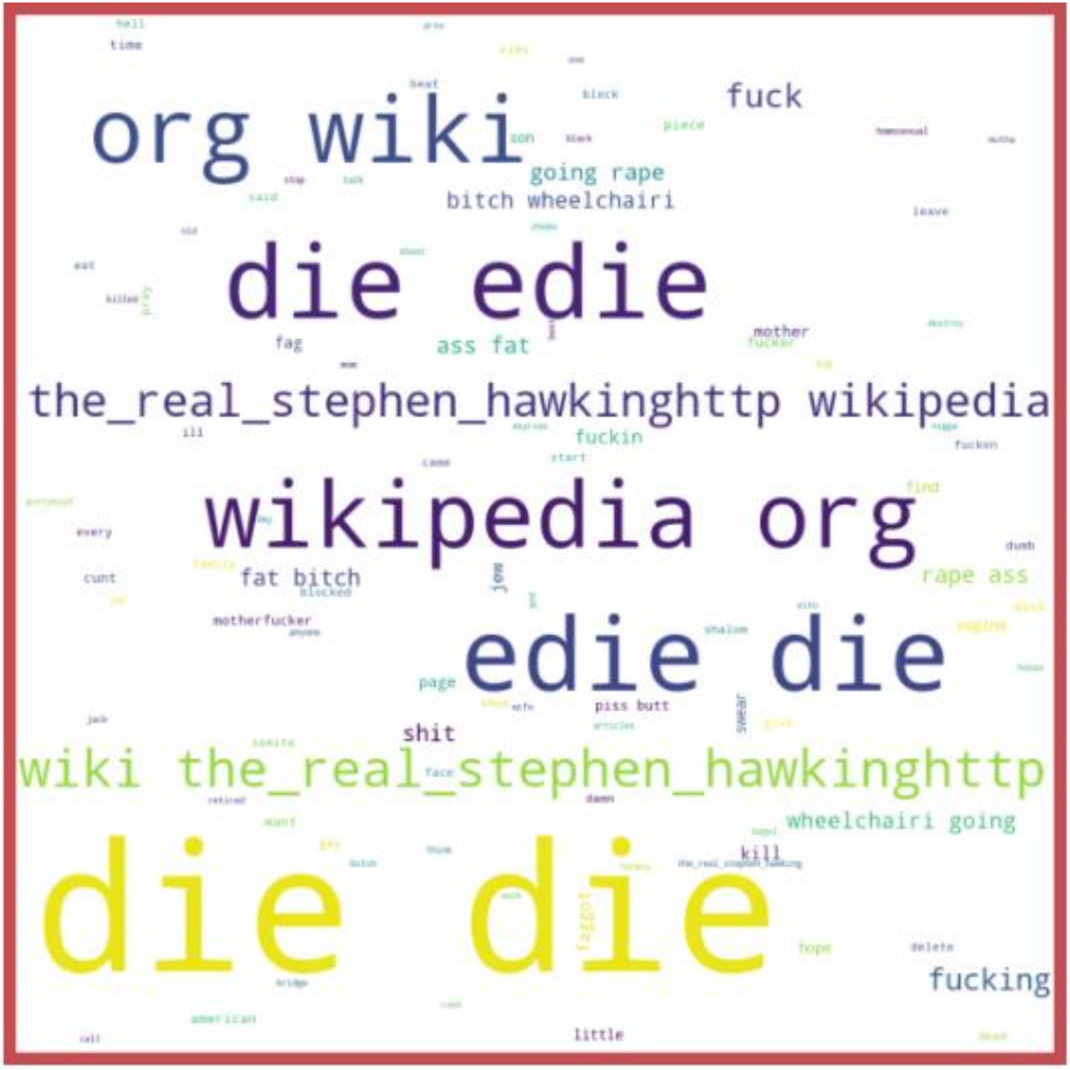
Abuse



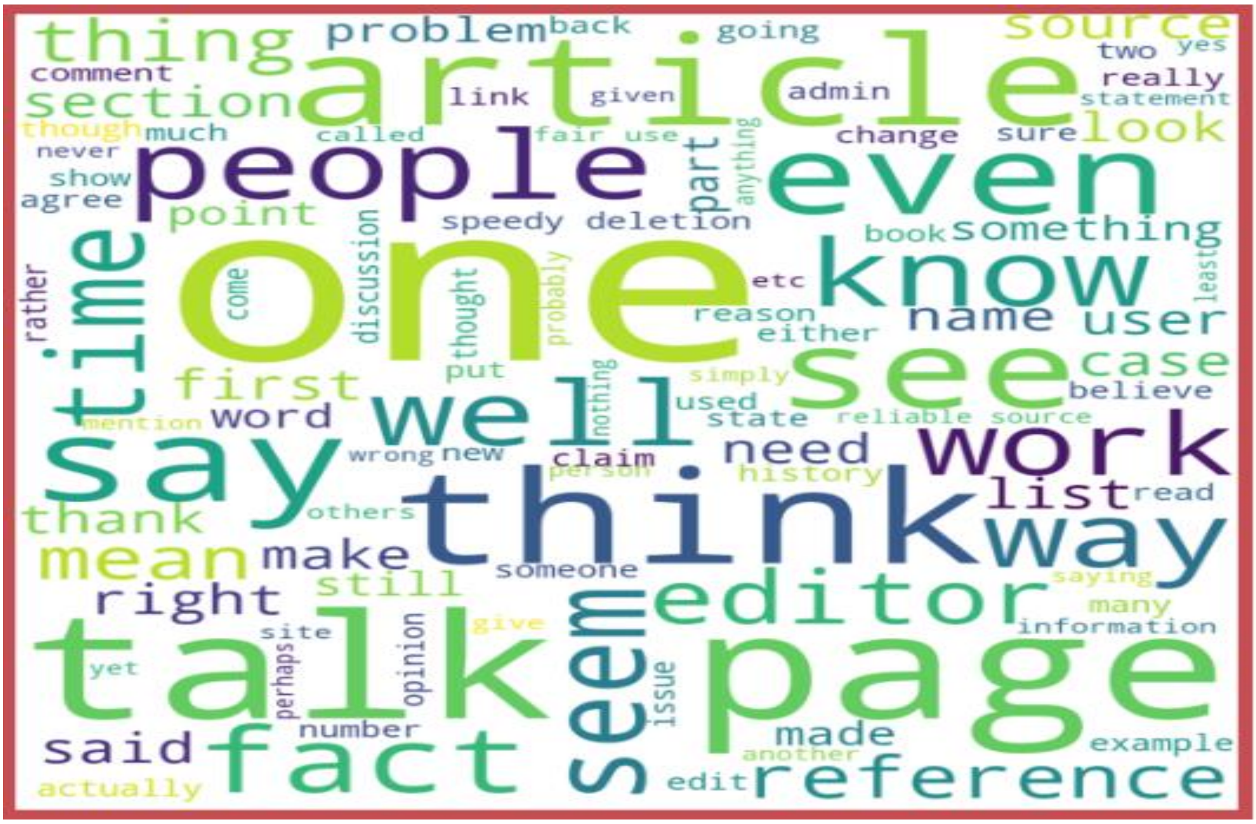
Loathe



"malignant", "highly_malignant", "rude", "threat", "abuse","loathe"



Neutral(All Clean)



- Interpretation of the Results

  From the above we visualization we can see that there are multiple words which are categories as multiple labels and we can see most frequent words which were used are on multiple labels, and we can also see most frequent words which were labelled as neutral.

CONCLUSION

- Key Findings and Conclusions of the Study

  So, the key findings and conclusion we got from the whole analysis that there are few words which are focus on the same categories as the comment start going from bad to worse, and we keep them categories in multiple labels, as only because of this our Bidirectional LSTM using embedding worked so well, as on based on few words we can classify the whole comments.

- Learning Outcomes of the Study in respect of Data Science

  There where al lot of learning outcomes we were able to see that how efficient and less time consuming keras text processing library can clean and vectorizes the comments using TFIDF, and also wanted to use all the binary relevance and adaptation algorithm, but due to memory limitation had to comment out those algorithm as kernel was continues to die of both local machine and googlecolab even tried 15 different times eliminating one algo at a time, just will keep working on that as I am very curious and keen to see the results of all other methods even using Naïve Byes multinomialNB expecting a very good results, and have future work application of combining TF-IDF with sentiment features.