­

Malignant Comments Classification

Submitted by:

Shama Tanweer

Internship-12

**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.

Towarsdatascience.com

Datacamp.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.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’ssometimes 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**

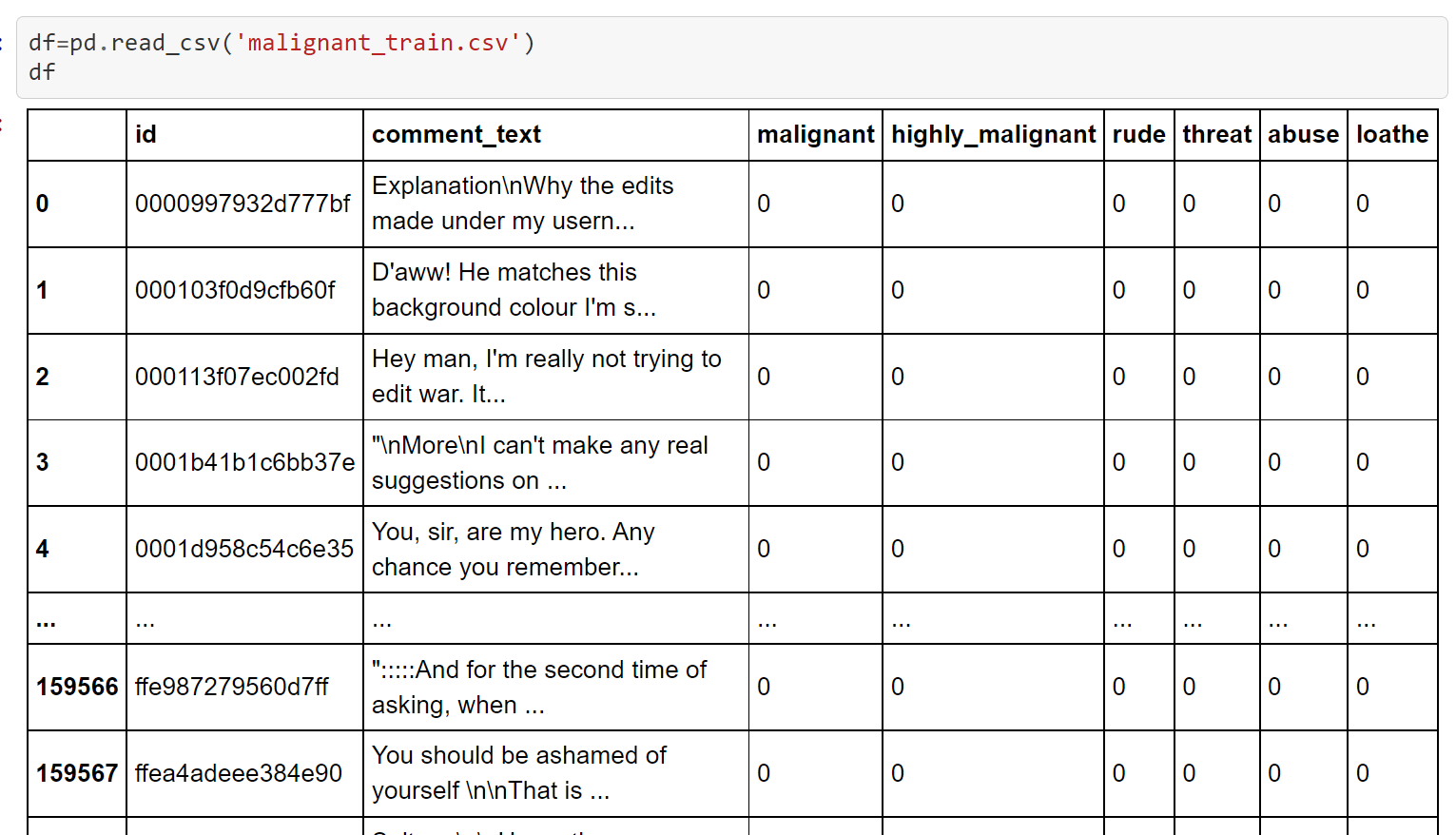
**Mathematical/Analytical modelling of problem**

* 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 attributes were of int type,except id & comment\_text(object type)
* Dataset did’nt contain any null value.

**Data source and their formats**

The data is provide to us from our clent database.The sample data is in .csv format

.The sample data for reference is shown below.



**DataSet description**

Train dataset samples contain 8 features 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.

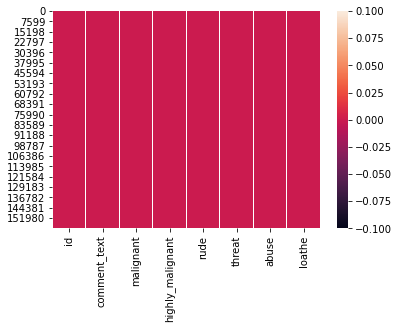
-**Comment text:** This column contains the comments extracted from various social media platforms.

* Data Pre-processing

Data usually comes from a variety of source & is often inconsistent,inaccurate.Data preprocessing helps to enhance the quality of dat and make it ready for various ML model.We have applied various methods for data preprocessing methods in this project .

1.checking for null values in each column:

sns.heatmap(df.isnull())



The above heatmap shows that no null values are pesent in the train dataset.

* Cleaning the raw data-It involves deletion of words or special characters that do not add meaning to the text.
* Important cleaning steps are:

1. Lowering case
2. Handling of special characters
3. Removal of stopwords
4. Handling of hyperlinks
5. Removing leading and trailing white space
6. Replacing urls with web address
7. Converted words to most suitable base form by using lemmatization

Adding new column “Comment\_length” which is the length of comment in number.

Adding new column “Clean\_comment” which is done by cleaning the column “comment-text”.

Adding one more column “comment\_clean\_length” which is count of length of comment column by mapping

* 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.

**Hardware and Software Requirements and Tools Used**

* 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.

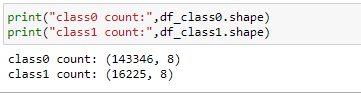
Library Used:

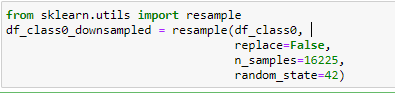
**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

**Understanding the problem is the first crucial steps in solving any problem.** From the given dataset it can be concluded that it is a MultiLabel Classification problem.Therefore I run my preprocessed data on 6 classification algorithm.

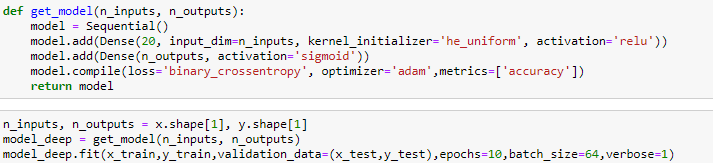
We also found that the dataset was highly imbalanced as class 1 had 16225, and class 0 had 143346, so we down sampled class 0 to 16225 to make the data balanced using resample.

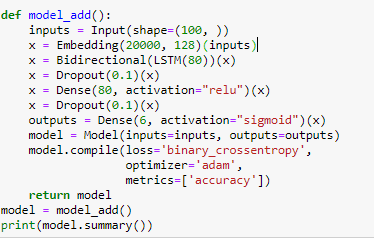




* Testing of Identified Approaches (Algorithms)

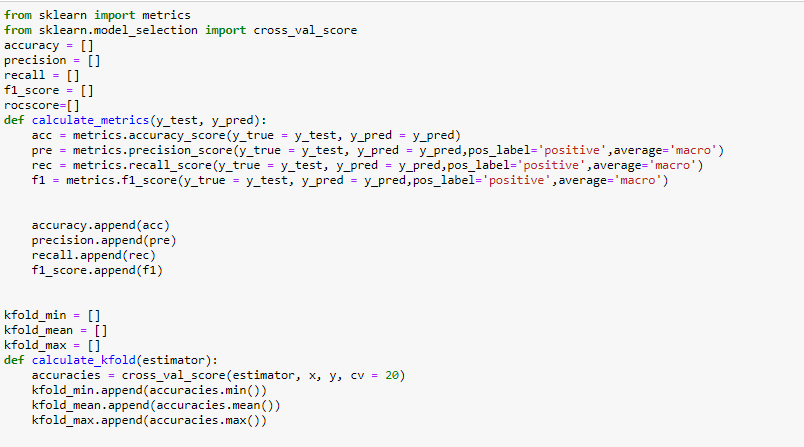




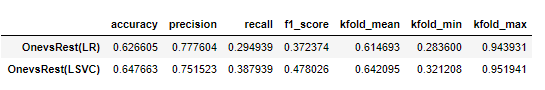


From the above we can see that the we have tried various methods and utilized pipelines in order to achieve a robust model, so we used OneVSRestclassifier on logistic regression and linear SVM, and used deep learning models and bidirectional LSTMwith embedding and achieve a better result.

* Key Metrics for success in solving problem under consideration



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.



* Visualizations

Malignanat



highly\_malignant



Rude



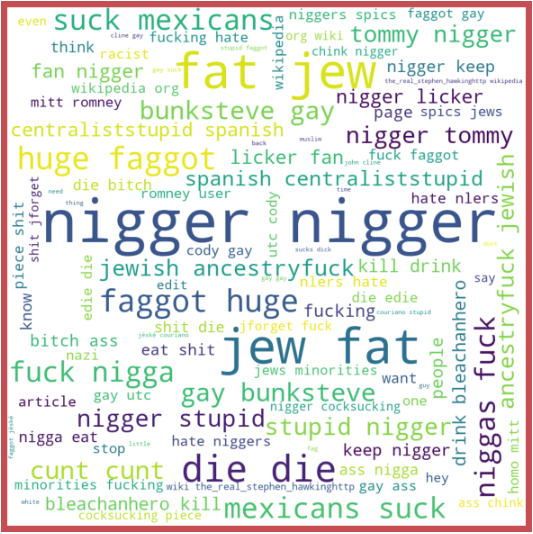
Threat



Abuse



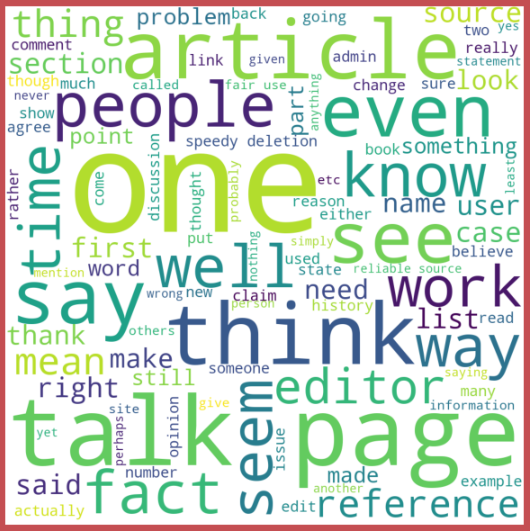
Loathe



"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 ByesmultinomialNB expecting a very good results, and have future work application of combining TF-IDF with sentiment features.