

***MALIGNANT COMMENTS CLASSIFICATION***

Submitted by:

Anu Sindhuja Galidevara

**ACKNOWLEDGMENT**

<https://www.researchgate.net/publication/338798595_Toxic_Comment_Detection_in_Online_Discussions>

<https://www.slideshare.net/ijtsrd/toxic-comment-classification-153346022>

<https://www.researchgate.net/publication/327345300_Challenges_for_Toxic_Comment_Classification_An_In-Depth_Error_Analysis>

<https://www.kaggle.com/adamschroeder/countvectorizer-tfidfvectorizer-predict-comments>

**INTRODUCTION**

Over a decade, social networking and social media have been growing in leaps and bounds. Today, people are able to express themselves and their opinions and discuss among others via these platforms. In such a scenario, it is obvious that debates may arise due to differences in opinion. But often these debates take a dirty side and may result in fights over the social media during which offensive language termed as toxic comments may be used from one side. These toxic comments may be threatening, obscene, insulting or identity-based hatred. So, these clearly pose the threat of abuse and harassment online.

Consequently, some people stop giving their opinions or give up seeking different opinions which result in unhealthy and unfair discussion. As a result, different platforms and communities find it very difficult to facilitate fair conversation and are often forced to either limit user comments or get dissolved by shutting down user comments completely. The Conversation AI team, a research group founded by Jigsaw and Google have been working on tools and techniques for providing an environment for healthy communication. They have also built publicly available models through the Perspective API on Comment Toxicity. But these models are sometimes prone to errors and does not provide the option to the users for choosing which type of toxicity, they are interested in finding. So, a more stable and versatile intelligent system is required for Toxic Comment Prevention in social communication.

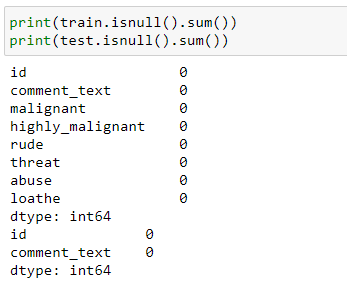
This model reads any piece of text (a text message or any comment appearing in social platform that can be toxic or nontoxic) and detects the type of toxicity it contains. The types of toxicity are simply toxic, severely toxic, obscene, threat, insult and identity-based hate. This overcomes the drawback of the model developed using Perspective API,showing all the types of toxicity contained in the comment.

**Analytical Problem Framing**

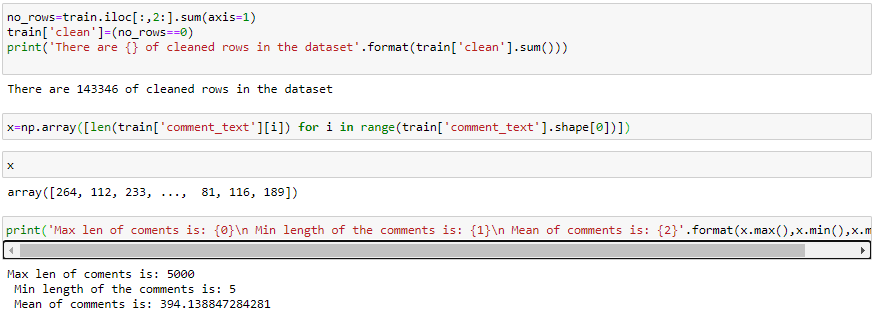
This document is structured as follows:

* First exploration: just to see what we have.
* Cleaning: time to make choices about unwanted data
* Feature engineering: time to be creative
* Result and lessons learned

1. As a first step, we must look if there are missing or anomalous data

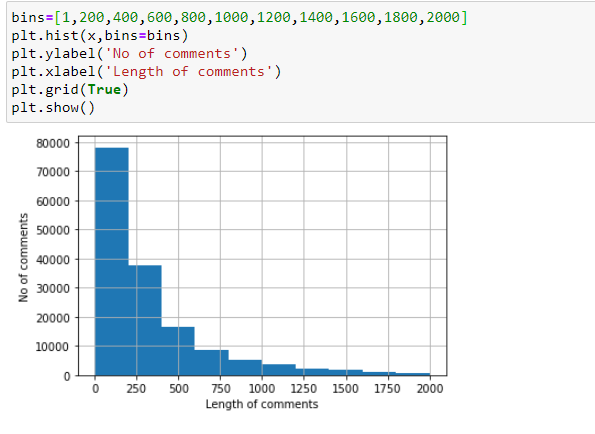


Take an insight of number of cleaned comments and length of comments

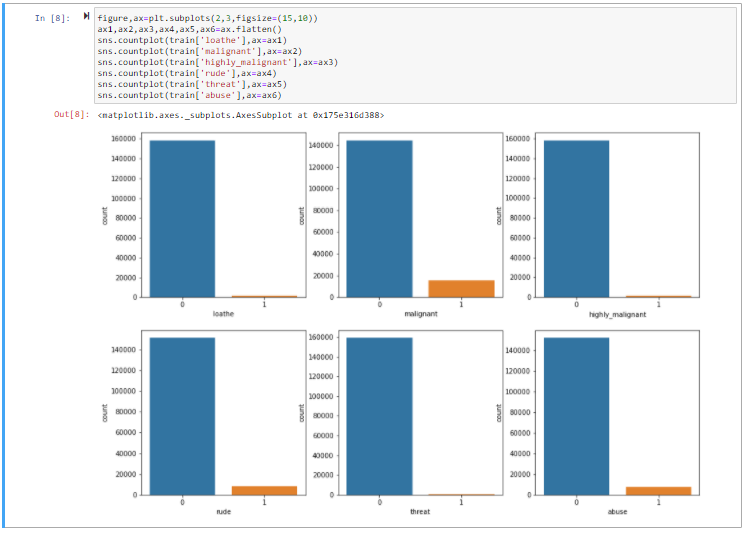


1. Data Visualization:

Visualizations are done in the form of Histograms showing the distribution of comment lengths over the whole corpus of the dataset



Count of different toxicity of comments in each label



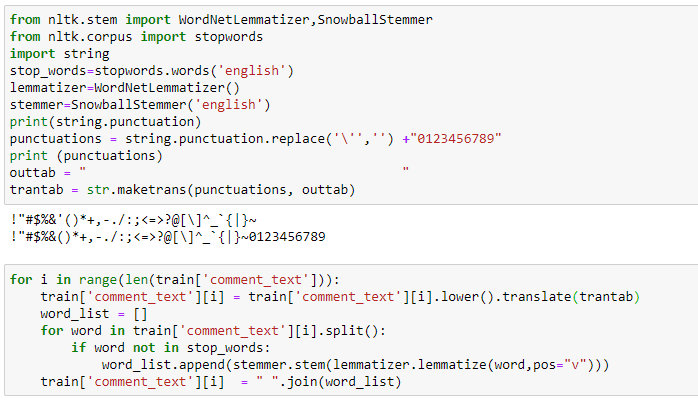
1. Text Preprocessing :

The text preprocessing techniques followed before processing the text data are:

- Removal of Punctuation: All the punctuation marks in every comment are removed.

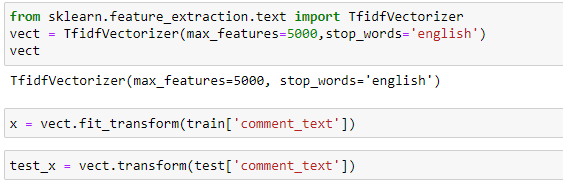
– Lemmatisation: Inflected forms of words which may be different verb forms or singular/plural forms etc. are called lemma. For ex. go and gone are inflected forms or lemma of the word, gone. The process of grouping these lemmas together is called Lemmatisation. So, Lemmatisation is performed for every comment.

– Removal of Stopwords: Frequently occurring common words like articles, prepositions etc. are called stopwords. So, stopwords are removed for each comment.



1. Tf-idf Transformer

The featured Bag-of-Words Model or Matrix for the whole corpus is transformed into a matrix whose every element is a product of Term Frequency (TF) and Inverse Document Frequency (IDF), combined as tf-idf. Term Frequency (TF) is defined as the ratio of the number of times a word or a term appears in a comment to the total number of words in the comment.



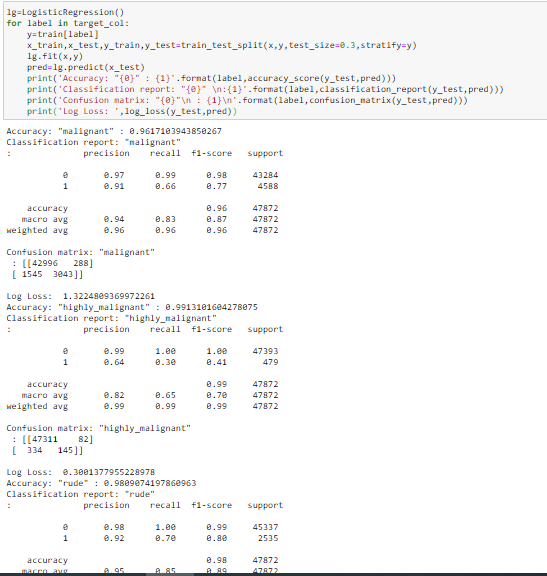
**Model/s Development and Evaluation**

The supervised machine learning can be further divided into two sub-groups: classification and regression. The problem with categorical outputs are grouped into classification problem, while the outputs of a regression problem are numerical. In this research, the output is either 0 or 1. The problem, hence, should be grouped into classification problem.

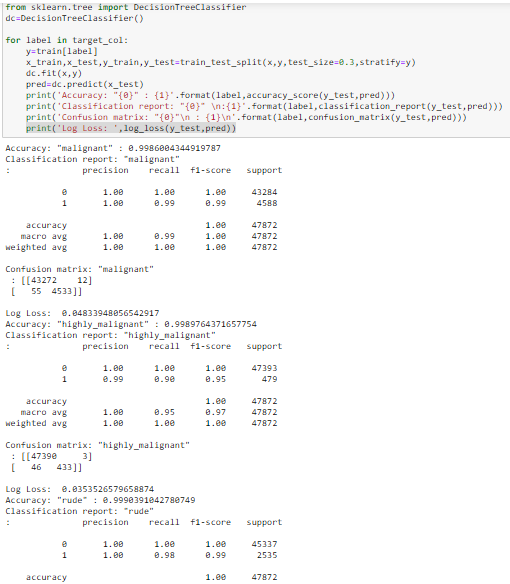
With text represented as vectors, we can finally train our model. Now we can actually use almost any sort of classification algorithms and see how these base models perform on the data by evaluating the metrics.

* Logistic Regression:

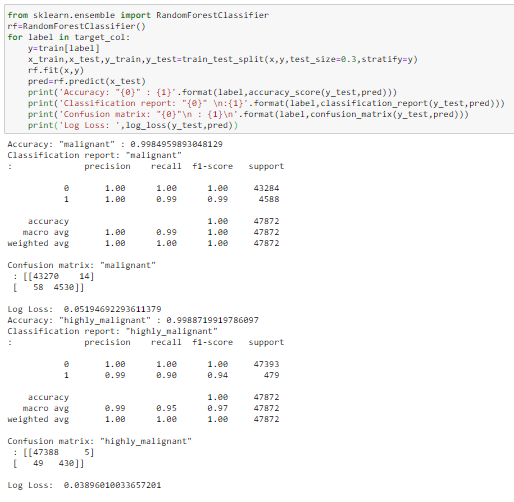
The logistic regression algorithm is based on the concept of probability of the predicted output which lies within 0 and 1 range:



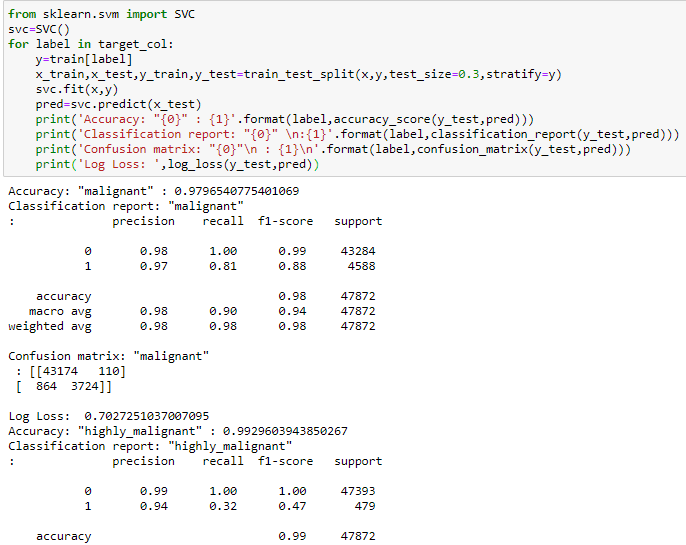
* Decision Tree Classifier:



* Random Forest Classifier:



* SVC:



**CONCLUSION**

In this document we proposed a Machine Learning for toxicity detection and its type identification in user comments. Finally, the Mean Validation Accuracy, so obtained, is 99% by Decision Tree Classifier which is far the highest ever numeric accuracy by any Comment Toxicity Detection Model. A more robust model can be developed by applying Grid Search Algorithm on the same dataset over the Machine Learning Algorithms for every pipeline, being used in order to obtain more better results and accurate classifications.