

Malignant Comments Classifier Project

Submitted by:

Vipin Kumar Singh

**ACKNOWLEDGMENT**

* Wikipedia : https://en.wikipedia.org/wiki/Multi-label\_classification
* Kaggle challenge page for datasets and ideas : <https://www.kaggle.com/c/>
* jigsaw-toxic-comment-classification-challenge
* Conversation AI git page : https://conversationai.github.io/
* Research Paper titled “Multi-label Classification: Problem Transformation methods -

classification” : <https://ac.els-cdn.com/>

* S1877050917319440/1-s2.0-S1877050917319440-main.pdf? \_tid=eced1a38-f8fa-11e7- b8ef-00000aab0f27&acdnat=1515914406\_0f244d3e6313bb049c435bf43 504bd52
* Research Paper titled “Benchmarking Multi-label Classification Algorithms” :
* <http://ceur-ws.org/Vol-1751/AICS_2016_paper_33.pdf>
* Problem Transformation Methods : https://www.analyticsvidhya.com/blog/ 2017/08/introduction-to-multi-label-classification/
* Research Paper on BP-MLL : http://citeseerx.ist.psu.edu/viewdoc/ download?doi=10.1.1.507.910&rep=rep1&type=pdf
* GridsearchCV on Sequential Models : https://dzubo.github.io/machinelearning/2017/05/25/increasing-model-accuracy-by-tuning-parameters.html

**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 aproblem 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 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.

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 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

* [Data exploration is the initial step in data analysis](https://www.sisense.com/blog/exploratory-data-analysis/), where users explore a large data set in an unstructured way to uncover initial patterns, characteristics, and points of interest. This process isn’t meant to reveal every bit of information a dataset holds, but rather to help create a broad picture of important trends and major points to study in greater detail.

Data exploration can use a combination of manual methods and automated tools such as [data visualizations](https://www.sisense.com/product/data-visualization/), charts, and initial reports.

* Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data. In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model to be studied.
* **Machine learning** (ML) is a type of artificial **intelligence** (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. **Machine learning** algorithms use historical data as input to predict new output values.

**Evaluation Metrics**

**→ Label bases metrics** include one-error, average precision, etc. These can be calculated for each labels, and then can be averaged for all without taking into account any relation between the labels if exists.

Average Precision (AP): Average precision is a measure that combines recall and precision for ranked retrieval results. For one information need, he average precision is the mean of the precision scores after each relevant document is retrieved, where, and are the precision and recall at the threshold.

**→ Example based metrics** include accuracy, hamming loss, etc. These are calculated for each of the examples and then averaged across the test set. Let –

Accuracy is defined as the proportion of correctly predicted labels to the total no. of labels for each instance.

Hamming-loss is defined as the symmetric difference between predicted and true labels, divided by the total no. of labels.

When we have a look at the data, we observe that every 1 in 10 samples is toxic, every 1 in 50 samples is obscene and insulting, but the occurrences of sample being severe\_toxic, threat and identity hate is extremely rare. Thus, we have skewed data, and accuracy as metric will not give us the required results. Thus, we will be using hamming-loss as the evaluation metric.

* Review of Literature

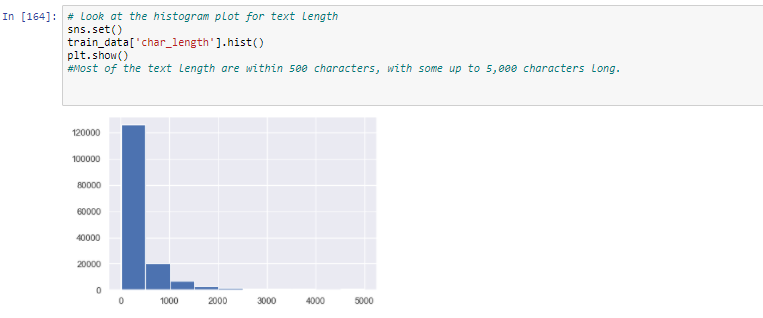
This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

We have to find the data mean, median, mod and also need to defined it mathematically. We also need to compare of one data to another using the mathatical appraoch to classify the data for useful purpose.

historgram: How the data is spreaded about its mean value which is calculated. It will show how the data is varying form one starting and end point.



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

* Data Sources and their formats

The project data is provided by the customer which is related to the online comment text. The data is stored in the tabular format with approximately 3,24,000 entries in it. The data basically orginated from a online application which store the user comment related to the specific social media website like facebook, youtube, twitter, etc.

The data set format description is given below in the table

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| **id** | A unique id aligned with each comment text. |
| **comment\_text** | It includes the comment text. |
| **malignant** | It is a column with binary values depicting which comments are malignant in nature. |
| **highly\_malignant** | Binary column with labels for highly malignant text. |
| **rude** | Binary column with labels for comments that are rude in nature. |
| **threat** | Binary column with labels for threatening context in the comments. |
| **abuse** | Binary column with labels with abusive behaviour. |
| **loathe** | Label to comments that are full of loathe and hatred. |

The intially data analysis is shows that the data set contains the user comment which can be classified in the below…

* Maliganant
* Highly\_magliganant
* Rude
* Threat
* Abuuse
* Loathe

There are total two dataset files are provided for the analysis purpose

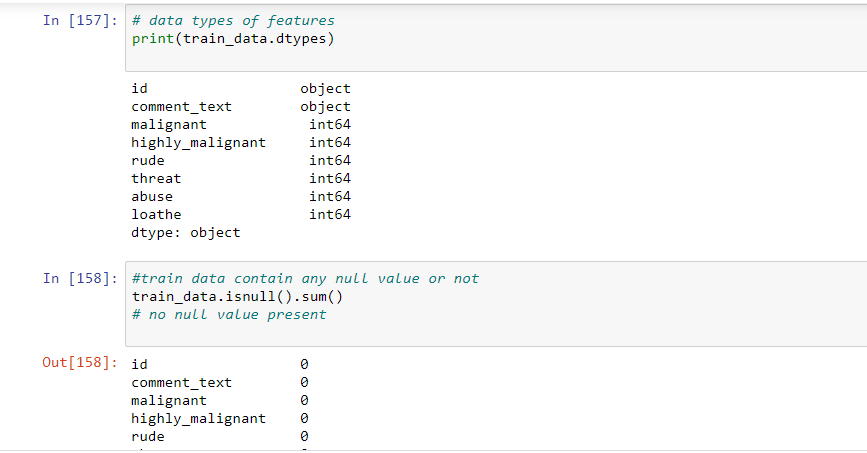
1. Train.csv – It contains the trainig set which contains comments with their binary labels.
2. Test.csv – The file contains the set for which the predictions are to be ddone. it inculdes the id and comment test.

* Data Preprocessing

**Data Exploration**

The dataset consists of the following fields.

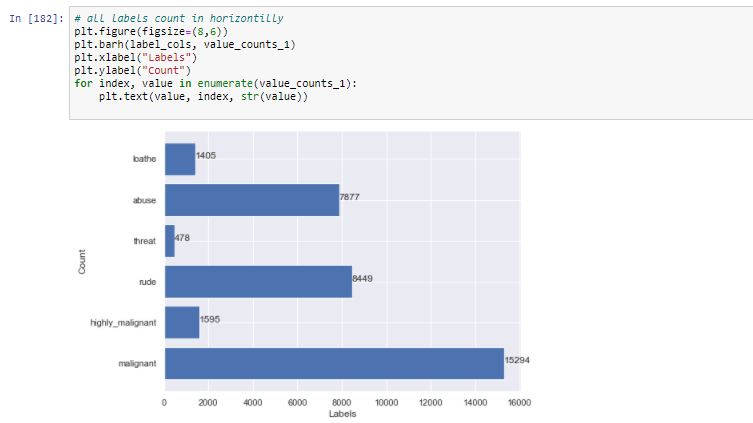
* id: An 8-diit integer value, to get the identify of the person who had written this comment
* comment\_text: A multi-line text field which contains the unfiltered comment
* maliganant: binary label which contains 0/1 (0 for no and 1 for yes)
* highly\_malgnant: binary label which contains 0/1
* rude: binary label which contains 0/1
* threat: binary label which contains 0/1
* abuse: binary label which contains 0/1
* loathe: binary label which contains 0/1



1. Data Preprocessing The following steps were taken to process the data: → A string without all punctuations to be prepared: 1. The string library contains punctuation characters. This is imported and all numbers are appended to this string. Our comment\_text field contains strings such as won't, didn't, etc. which contain apostrophe character('). To prevent these words from being converted to wont or didn't, the character ' represented as \' in escape sequence notation is replaced by empty character in the punctuation string. 2. make\_trans( intab, outtab) function is used. It returns a translation table that maps each character in the intab into the character at the same position in the outtab
2. Stop words are those words that are frequently used in both written and verbal communication and thereby do not have either a positive or negative impact on our statement like “is, this, us, etc.”.
3. Single letter words if existing or created due to any preprocessing step do not convey any useful meaning and so they can be directly removed. Hence letters from b to z, will be added to the list of stop words imported directly.
4. Stemming and Lemmatizing:

* The process of converting inflected/derived words to their word stem or the root form is called stemming. Many similar origin words are converted to the same word e.g. words like "stems", "stemmer", "stemming", "stemmed" as based on "stem".
* Lemmatizing is the process of grouping together the inflected forms of a word so they can be analyzed as a single item. This is quite similar to stemming in its working but differs since it depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document.
* The wordnet library in nltk will be used for this purpose. Stemmer and Lemmatizer are also imported from nltk.

1. Applying Count Vectorizer:
2. To convert a string of words into a matrix of words with column headers represented by words and their values signifying the frequency of occurrence of the word Count Vectorizer is used.
3. Stop words were accepted, convert to lowercase, and regular expression as its parameters. Here, we will be supplying our custom list of stop words created earlier and using lowercase option. Regular expression will have its default value.
4. Splitting dataset into Training and Testing:
5. Since the system was going out of memory using train\_test\_split, I had jumbled all the indexes in the beginning itself.
6. The shuffle function defined here performs the task of assigning first 2/3rd values to train and remaining 1/3rd values to the test set.





**Text Preprocessing**

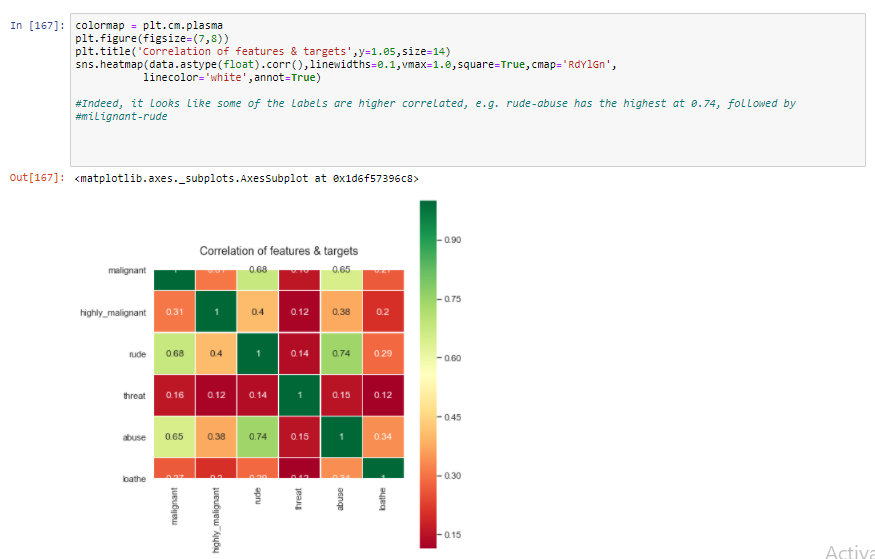
The text preprocessing techniques followed before processing the text data are:

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

**– Lemmatisation:** Inflected forms ofwordswhichmay be different verb forms or sigular/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 lemma together is called Lemmatisation. So, Lemmatisation is performed for every comment.

**– RemovalofStopwords:**Frequently occurring commonwordslike articles, prepositions etc. are called stopwords. So, stopwords are removed for each comment.

* Data Inputs- Logic- Output Relationships
* Correlation: It is a method that define the relationship between the features. Means if one feature is increaseing what will be the affect of another.



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

For this project we have not taken any assummption

* Hardware and Software Requirements and Tools Used

Hardware Requirement

The data classification and relation calculation require the hardware resources to be utilized in order to get the requried result. For this project the list of hardware resources required are given below in the table.

Recommended hardware required

**OS:** Windows 10 64 Bit (latest Service Pack)  
**CPU:**Intel Core i5-2500K or AMD Ryzen R5 1600X  
**RAM:** 12GB RAM  
**HDD (at launch\*):** 82GB HD space  
**Video:** NVIDIA GeForce GTX 970 / GTX 1660 Super or Radeon R9 390 / AMD RX 580

Minimum hardware requried

**OS:**Windows 7 64-Bit (SP1) or Windows 10 64-Bit (1803 or later)  
**CPU:** Intel Core i3-4340 or AMD FX-6300  
**RAM:** 8GB RAM  
**HDD (at launch\*):** Multiplayer only: 35GB HD space / All game modes: 82GB HD space  
**Video:** NVIDIA GeForce GTX 670 / GeForce GTX 1650 or AMD Radeon HD 7950

Software Requirement

The list of software tool which is required to be used in this data classification are given below for different categories.

* Programming Language
  + Python – Version 3.7
  + R
* Database
  + MySQL – version 8.1
* Software Tool
  + Worksheet software – MS Excel
  + Anaconda
  + Jyptyer Notebook

**Model/s Development and Evaluation**

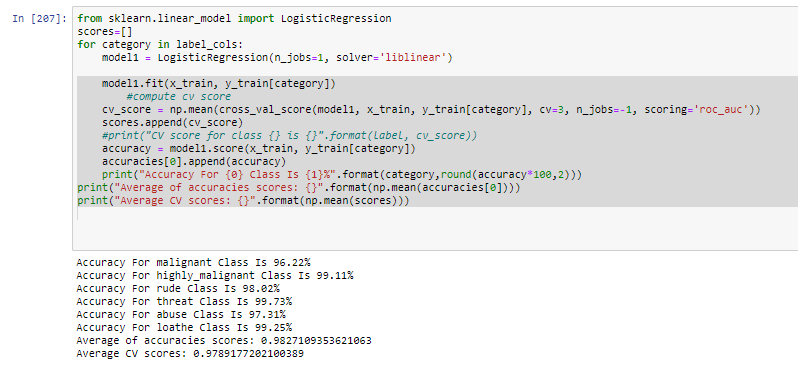
* Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

* Run and Evaluate selected models

List of all the algorithm used for the data classificaiton and analysis are given below.

* Logistic Regression:



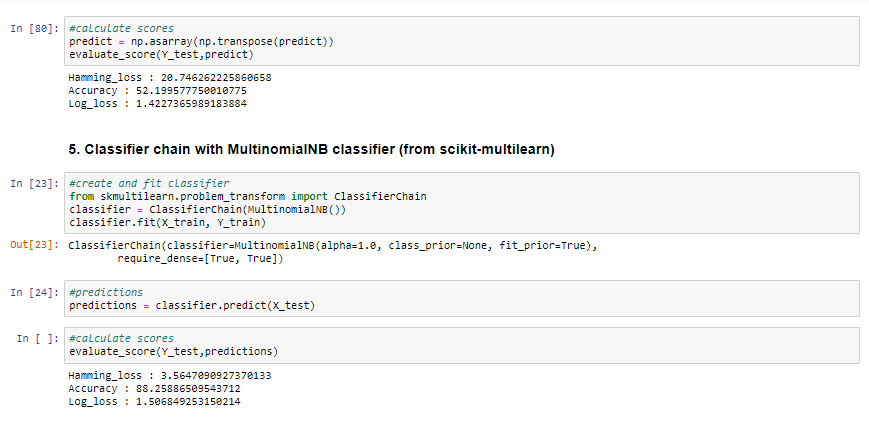
* Linear Support Vector Classifier:



* Multinominal Naive bayes



* Key Metrics for success in solving problem under consideration

 Interpretation of the Results

Model efficiency is near about 88%.

**CONCLUSION**

* Learning Outcomes of the Study in respect of Data Science

→The first step involved collecting data and deciding what part of it is suitable for training : This step was extremely crucial since including only very small length comments would give poor results if the length was increased whereas including very long length comments would increase the number of words drastically, hence increasing the training time exponentially and causing system (jupyter kernel) to go out of memory and die eventually.

→ The second major step was performing cleaning of data including punctuation removal, stop word removal, stemming and lemmatizing: This step was also crucial since the occurrence of similar origin words but having different spellings will intend to give similar classification, but computer cannot recognize this on its own. Hence, this step helped to a large extent in both removing and modifying existing words.

→ The third step was choosing models to train on: Since I had a wide variety of models( 3 for problem transformation) and classifiers(not bounded) along with number of adaptation models in BP-MLL, selecting which all models to train and test took lots of efforts.

* Limitations of this work and Scope for Future Work
* Limitation

→Although we have tried quite several parameters in refining my model, there can exist a better model which gives greater accuracy.

→ Yes. We were unable to find a clear implementation of the Adaboost.MH decision tree model which we had planned to use. The scikit-multilearn library doesn’t even mention of such a model. Also, the research papers were a little vague regarding implementation details.

* Future Scope

The current project predicts the type or toxicity in the comment. We are planning to add the following features in the future:

→ Analyse which age group is being toxic towards a particular group or brand.

→ Add feature to automatically sensitize words which are classified as toxic.

→ Automatically send alerts to the concerned authority if threats are classified as severe.

→ Build a feedback loop to further increase the efficiency of the model.

→ Handle mistakes and short forms of words to get better accuracy of the result.