

**Malignant Comments Classifier**

**Submitted by: Harshita Panchamia**

**Batch No: Internship 25**

**ACKNOWLEDGMENT**

For this particular task, I referred the following websites and articles when stuck,

<https://www.kaggle.com/surekharamireddy/malignant-comment-classification>

<https://monkeylearn.com/blog/data-preprocessing/>

<https://towardsdatascience.com/text-normalization-for-natural-language-processing-nlp-70a314bfa646>

<https://towardsdatascience.com/tf-idf-simplified-aba19d5f5530>

**INTRODUCTION**

Online forums and social media platforms have provided individuals with the means to put forward their thoughts and freely express their opinion on various issues and incidents. In some cases, these online comments contain explicit language which may hurt the readers. Comments containing explicit language can be classified into myriad categories such as Toxic, Severe Toxic, Obscene, Threat, Insult, and Identity Hate. The threat of abuse and harassment means that many people stop expressing themselves and give up on seeking different opinions.

To protect users from being exposed to offensive language on online forums or social media sites, companies have started flagging comments and blocking users who are found guilty of using unpleasant language. Several Machine Learning models have been developed and deployed to filter out the unruly language and protect internet users from becoming victims of online harassment and cyberbullying.

* **Conceptual Background of the Domain Problem**

The main purpose of building this model is to prevent the abusive comment which in turn will Detroit the mindset of an individual or people, now-a-days a lot of abusive and lethargic comment can be seen on various social media platform which create a negative environment among the people and community, so to stop this type of activity a machine learning model is built to identify the malignant text and filter it out as soon as it encounters it.

* **Review of Literature**

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 Using Machine Learning Techniques. In this we are investigating the application of supervised machine learning techniques to predict the comments. The predictions are based on historical data collected from websites like twitter etc. Different techniques. To build a model for predicting the comments we have used Supervised machine learning.

* **Motivation for the Problem Undertaken**

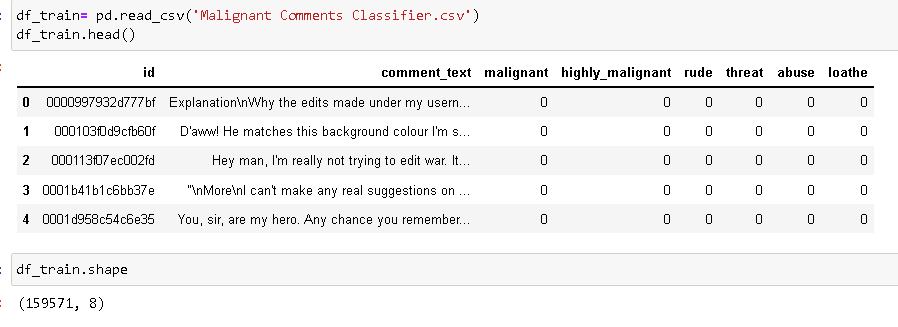
Online platforms when used by normal people can only be comfortably used by them only when they feel that they can express themselves freely and without any reluctance. If they come across any kind of a malignant or toxic type of a reply which can also be a threat or an insult or any kind of harassment which makes them uncomfortable, they might defer to use the social media platform in future. Thus, it becomes extremely essential for any organization or community to have an automated system which can efficiently identify and keep a track of all such comments and thus take any respective action for it, such as reporting or blocking the same to prevent any such kind of issues in the future.

* **Dataset Description**

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’. There are various comments which have multiple labels. The first attribute is a unique ID associated with each comment. The label can be either 0 or 1, where 0 denotes a NO while 1 denotes a YES.

The data set includes:

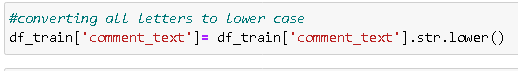
* id **:** person who have written the comment is generalised by id.
* comment\_text **:** thoughts of person.
* malignant **:** binary label which contains 0/1.
* highly-malignant: binary label which contains 0/1.
* rude: binary label which contains 0/1.
* loathe**:** binary label which contains 0/1.
* abuse : binary label which contains 0/1.
* threat : binary label which contains 0/1



**Data Preprocessing**

Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning. Raw, real-world data in the form of text, images, video, etc., is messy. Not only may it contain errors and inconsistencies, but it is often incomplete, and doesn’t have a regular, uniform design. Machines like to process nice and tidy information – they read data as 1s and 0s. So calculating structured data, like whole numbers and percentages is easy. However, [unstructured data](https://monkeylearn.com/blog/structured-data-vs-unstructured-data/), in the form of text and images must first be cleaned and formatted before analysis.

* **Converting upper case to lower case**: It converts all the upper-case text in the comment to lower case, it is an important technique as it helps in cleaning the data.

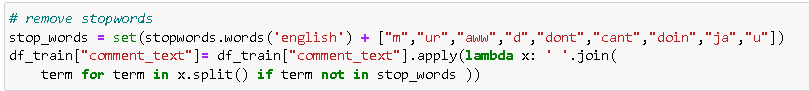


* **Text Normalisation**: As I was now certain that there are no missing records in my data, I decided to start with data pre-processing. Firstly, I decided to normalize the text data since comments from online forums usually contain inconsistent language, use of special characters in place of letters (e.g., @rgument), as well as the use of numbers to represent letters (e.g., n0t). To tackle such inconsistencies in data, I decided to use Regex***.*** The text normalization steps that I performed are listed below: -
* Removing Characters in between Text.
* Removing Repeated Characters.
* Converting data to lower-case.
* Removing Punctuation.
* Removing unnecessary white spaces in between words.
* Removing “\n”.
* Removing Non-English characters.





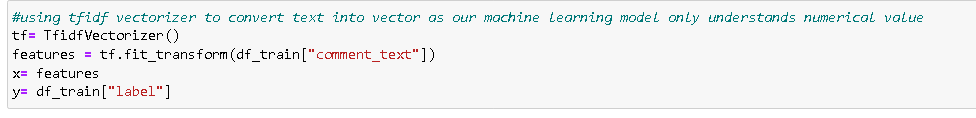
* **Stop word Removal:** Stop words are those words that are frequently used in both written and verbal communication and thereby do not have either a positive/negative impact on our statement. Stop words are a set of commonly used words in a language. Examples of stop words in English are “a”, “the”, “is”, “are” and etc. The intuition behind using stop words is that, by removing low information words from text, we can focus on the important words instead.

****

* Lemmatisation: lemmatisation is the process of grouping together of different inflated form words so they can be analysed as a single item.

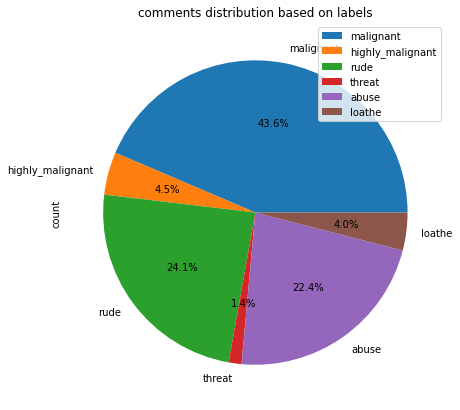
****

* Tf-Idfvectorization **:** Term frequency-inverse document frequency is a text vectorizer that transforms the text into a usable vector. It combines 2 concepts, Term Frequency (TF) and Document Frequency (DF). The term frequency is the number of occurrences of a specific term in a document. Term frequency indicates how important a specific term in a document. Term frequency represents every text from the data as a matrix whose rows are the number of documents and columns are the number of distinct terms throughout all documents. Document frequency is the number of documents containing a specific term. Document frequency indicates how common the term is.

****

**Data Visualizations**

According to the data, malignant comments has highest percentage of 43.6% followed by rude 24.1%, abuse 22.4%, highly malignant 4.5%, loathe 4% and threat 1.4% .



**Model Building**

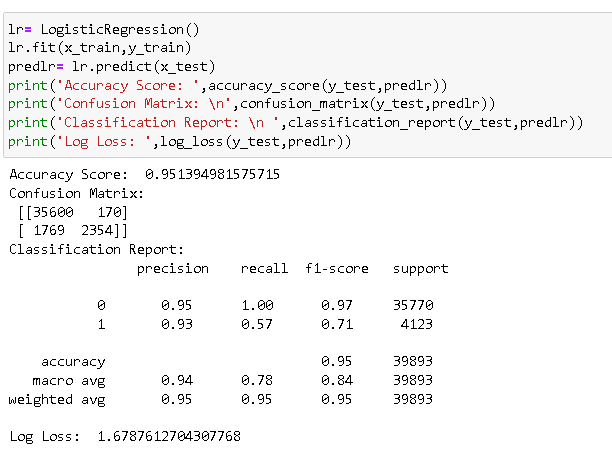
Train-Testsplit: There are two primary phases in the system:

1. Training phase: The system is trained by using the data in the data set and fits a model (line/curve) based on the algorithm chosen accordingly.

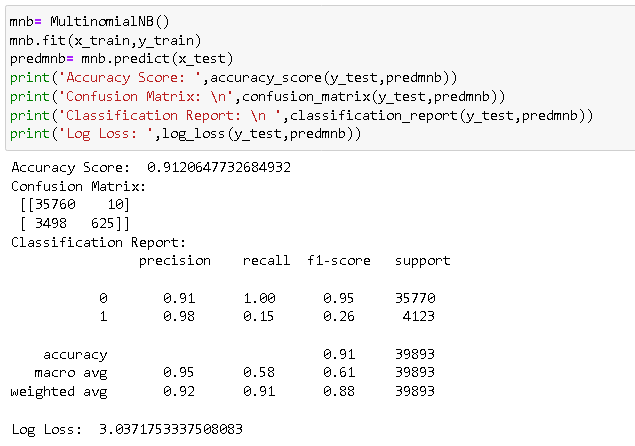
2. Testing phase: the system is provided with the testing data, and it is tested for its working. The accuracy is checked. And therefore, the data that is used to train the model or test it, must be appropriate. The system is designed to detect and predict price of used car and hence appropriate algorithms must be used to do the two different tasks.

I use 4 different algorithms for model building:

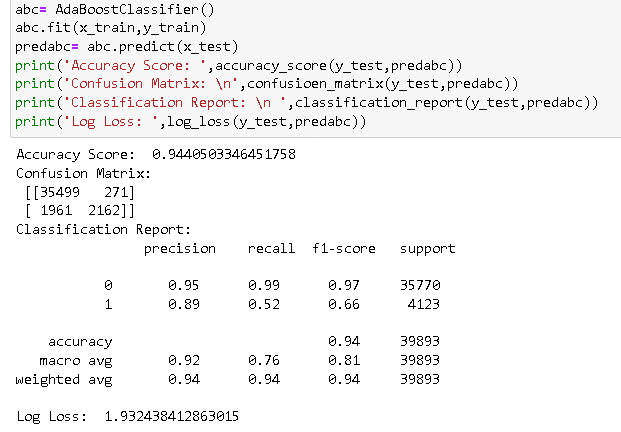
1. Logistic Regression



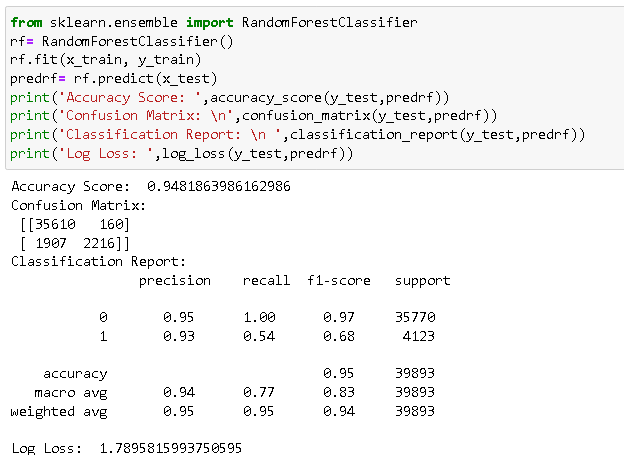
1. Multinomial NB



1. AdaBoost Classifier

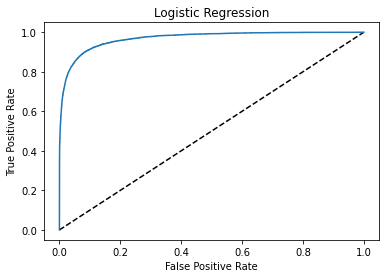


1. Random Forest Classifier

****

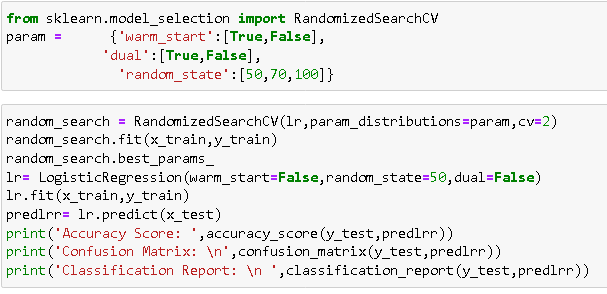
AUC ROC Curve:

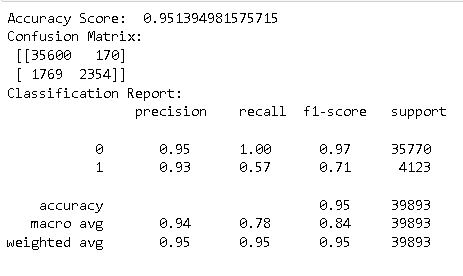
AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.



**Hyper parameter tunning**

Many machine learning algorithms are used to predict. However, the prediction accuracy of these algorithms depends heavily on the given data when training the model. If the data is in bad shape, the model will be overfitted and inefficient, which means that data pre-processing is an important part of this experiment and will affect the final results. Thus, multiple combinations of pre-processing methods need to be tested before getting the data ready to be used in training. After analyzing every model logistic regression shows good accuracy and cv with least difference and on doing hyper parameter tuning it accuracy reaches to 95%.





Conclusion

This research work focuses on developing a model that would automatically classify a comment as either malignant or non-malignant using logistic regression. Therefore, this study aims to develop a multi-headed model to detect different types of malignant comment like threats, rude, abusive, and loathe. By collecting and preprocessing malignant comments for training and testing using term frequency- inverse document frequency (TF-IDF) algorithm, developing a multi-headed model will detect different types of malignant comment using logistic regression to train the dataset, and evaluate the model using confusion metrics