

**Malignant Comments Classification Project**

Submitted by:

N THARANI

**ACKNOWLEDGMENT**

Various research papers and websites are used. Research papers like [An Automated Toxicity Classification on Social Media](https://link.springer.com/article/10.1057/s41264-020-00085-7#auth-K__Ramanathan-Kalimuthu) paper by [Ahmad Alsharef](https://www.ncbi.nlm.nih.gov/pubmed/?term=Alsharef%20A%5BAuthor%5D&cauthor=true&cauthor_uid=35211168), [Karan Aggarwal](https://www.ncbi.nlm.nih.gov/pubmed/?term=Aggarwal%20K%5BAuthor%5D&cauthor=true&cauthor_uid=35211168), [Sonia](https://www.ncbi.nlm.nih.gov/pubmed/?term=Sonia%20%5BAuthor%5D&cauthor=true&cauthor_uid=35211168), [Deepika Koundal](https://www.ncbi.nlm.nih.gov/pubmed/?term=Koundal%20D%5BAuthor%5D&cauthor=true&cauthor_uid=35211168), [Hashem Alyami](https://www.ncbi.nlm.nih.gov/pubmed/?term=Alyami%20H%5BAuthor%5D&cauthor=true&cauthor_uid=35211168) and [Darine Ameyed](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ameyed%20D%5BAuthor%5D&cauthor=true&cauthor_uid=35211168); A New Application of Social Impact in Social Media for Overcoming Fake News paper by [Cristina M. Pulido](https://www.ncbi.nlm.nih.gov/pubmed/?term=Pulido%20CM%5BAuthor%5D&cauthor=true&cauthor_uid=32260048), [Laura Ruiz-Eugenio](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ruiz-Eugenio%20L%5BAuthor%5D&cauthor=true&cauthor_uid=32260048), [Gisela Redondo-Sama](https://www.ncbi.nlm.nih.gov/pubmed/?term=Redondo-Sama%20G%5BAuthor%5D&cauthor=true&cauthor_uid=32260048) and [Beatriz Villarejo-Carballido](https://www.ncbi.nlm.nih.gov/pubmed/?term=Villarejo-Carballido%20B%5BAuthor%5D&cauthor=true&cauthor_uid=32260048); How Malicious Comments on Social Media Has a Huge Impact on the Negative Attitude paper by Puiyu Huang; How Vicious Do You Think Your Social Media Comments Are? article by [Navjot Bians](https://medium.com/@biansnavjot?source=post_page-----204128141eab--------------------------------); Why People Post Benevolent and Malicious Comments Online paper by So-Hyun Lee, Hee-Woong Kim, etc., are used. Website like researchgate.net, geeksforgeeks, etc., are used as references. The data is received from the client which is their own data.

Thanking SWATI MAHASETH, my guide from FLIPROBO TECHNOLOGIES for clearing all my doubts while undergoing the project.

**INTRODUCTION**

* Business Problem

With the proliferation of smart devices and mobile and social network environments, the social side effects of these technologies, including cyberbullying through malicious comments and rumors, have become more serious. Malicious online comments have emerged as an unwelcome social issue worldwide.

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

In this age of the Internet, using social software has become an indispensable part of our lives. Although social software allows us to share our daily life, there will always be negative comments on social media. Since these comments are very common on social media, some people may think that these bad comments will only make the publisher’s mood down for a moment. In fact, this small factor may cause some bigger changes to the user’s mind like the butterfly effect.

It has become evident that [human behaviour is changing](https://scholarcommons.scu.edu/engl_176/2/); our emotions are getting attached to the likes, comments and tags we receive on social media. We get both good and bad comments but seeing hateful words, slurs and harmful ideas on digital platforms on a daily basis make it look normal when it shouldn’t be. The impact of toxic comments is much more catastrophic than we think. It not only hurts one’s self-esteem or deters people from having meaningful discussions, but also provokes people to such sinister acts as recent capital riots at US Congress and attacks on farmers for protesting in India. Therefore, having a solid toxicity flagging system in place is important if we want to maintain a civilized environment on social media platforms to effectively facilitate conversations.

* 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.

Certain websites and papers that helped me to take insights from are:

1. [An Automated Toxicity Classification on Social Media](https://link.springer.com/article/10.1057/s41264-020-00085-7" \l "auth-K__Ramanathan-Kalimuthu) paper by [Ahmad Alsharef](https://www.ncbi.nlm.nih.gov/pubmed/?term=Alsharef%20A%5BAuthor%5D&cauthor=true&cauthor_uid=35211168), [Karan Aggarwal](https://www.ncbi.nlm.nih.gov/pubmed/?term=Aggarwal%20K%5BAuthor%5D&cauthor=true&cauthor_uid=35211168), [Sonia](https://www.ncbi.nlm.nih.gov/pubmed/?term=Sonia%20%5BAuthor%5D&cauthor=true&cauthor_uid=35211168), [Deepika Koundal](https://www.ncbi.nlm.nih.gov/pubmed/?term=Koundal%20D%5BAuthor%5D&cauthor=true&cauthor_uid=35211168), [Hashem Alyami](https://www.ncbi.nlm.nih.gov/pubmed/?term=Alyami%20H%5BAuthor%5D&cauthor=true&cauthor_uid=35211168) and [Darine Ameyed](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ameyed%20D%5BAuthor%5D&cauthor=true&cauthor_uid=35211168)
2. A New Application of Social Impact in Social Media for Overcoming Fake News paper by [Cristina M. Pulido](https://www.ncbi.nlm.nih.gov/pubmed/?term=Pulido%20CM%5BAuthor%5D&cauthor=true&cauthor_uid=32260048), [Laura Ruiz-Eugenio](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ruiz-Eugenio%20L%5BAuthor%5D&cauthor=true&cauthor_uid=32260048), [Gisela Redondo-Sama](https://www.ncbi.nlm.nih.gov/pubmed/?term=Redondo-Sama%20G%5BAuthor%5D&cauthor=true&cauthor_uid=32260048) and [Beatriz Villarejo-Carballido](https://www.ncbi.nlm.nih.gov/pubmed/?term=Villarejo-Carballido%20B%5BAuthor%5D&cauthor=true&cauthor_uid=32260048)
3. How Malicious Comments on Social Media Has a Huge Impact on the Negative Attitude paper by Puiyu Huang
4. How Vicious Do You Think Your Social Media Comments Are? article by [Navjot Bians](https://medium.com/@biansnavjot?source=post_page-----204128141eab--------------------------------)
5. Why People Post Benevolent and Malicious Comments Online paper by So-Hyun Lee, Hee-Woong Kim

* Motivation for the Problem Undertaken

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.

This analysis contributes to identifying how interactions in these forms of social media depend on the type of information shared. The results indicate that messages focused on comments that are mostly aggressive, those based on evidence of social impact are respectful and transformative, and finally, deliberation contexts promoted in social media overcome false information.

The main motivation of this project is that the malignant comments can be classified as yes/no and it can be controlled using this model to screen prospective comments.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

[Natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing) (NLP) is a field that focuses on making natural human language usable by computer programs. **NLTK**, or [Natural Language Toolkit](https://www.nltk.org/), is a Python package that you can use for NLP. A lot of the data that could be analyzing is [unstructured data](https://en.wikipedia.org/wiki/Unstructured_data) and contains human-readable text. Before analyzing that data programmatically, first need to preprocess it. Various statistical, mathematical, analytical algorithms are used. Experimental design, etc are done for the problem.

* Data Sources and their formats

The data is received from the client. The data is in excel format which can be imported using pandas from local library easily.

* Data Preprocessing Done

Dropped id column which is of no use in building the model. Checked for null values where there is none present. The dataset doesn't have any duplicates. No correlation found and skewness found. Checking the balance of the dataset and length of strings present in each row. Convert all messages to lower case, replace email addresses with 'email', replace URLs with 'webaddress', replace money symbols with 'moneysymb', replace 10 digit phone numbers (formats include paranthesis, spaces, no spaces, dashes) with 'phonenumber', replace numbers with 'numbr'. After cleaning the dataset, checked the length of string in each row. Target columns are separated and transform the data into vectors only which will be considered by the model.

* Data Inputs- Logic- Output Relationships

There are various columns used to detect the comments whether those are malignant or not. All the features which are helping to detect the malignant comments is visualised using matplotlib and seaborn. The relationship between the features are determined. After data cleaning, pre-processing, model is built.

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

In this age of the Internet, there has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Only after checking out the model after building, online hate and abuse comment classifier can be found.

* Hardware and Software Requirements and Tools Used

Any laptop and computer can be used as hardware. Processor used is Intel(R) Core(TM) i7-4510U CPU. System type is 64-bit operating system, x-64 based processor. RAM of the systerm is 8.00 GB. Microsoft Windows 8.1 version 6.3 is the OS used. Python 2.7.10 is used with the interface Jupyter notebook with many installed libraries.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

There are statistical and analytical approaches in problem-solving. Data cleaning, data pre-processing, comes under statistical approach whereas data manipulation, vectorizing, creating ML models comes under analytical approach.

**Statistical:**

* + - ***Experimental Design*** - Methods to design systematic experiments to compare the effect of independent variables on an outcome
  + Univariate - measure relies on only one variable - Statistical summary, information on variables, etc
  + Bivariate - measure relies on two variables - Correlation, multicollinearity
    - ***WordNetLemmatizer*** - used to convert the similar words to a meaning word which is equally and similar to it
    - ***stopwords*** - Used to detect the words which does not help in determining the language or determining the result of the model
    - ***Re-sampling Methods*** - Train Test Split is used to systematically split a dataset into subsets for the purposes of training and evaluating this predictive model
    - ***Statistical Hypothesis Tests*** - Cross validation quantifies the likelihood of observing the result given an assumption or expectation about the result whether the model is overfitting/underfitting or fitting good
    - ***Estimation Statistics*** - GridSearchCV is used to quantify the best parameter from the listed to fit in the model and give better result. It uses data analysis framework which has a combination of effect sizes, confidence intervals, precision planning, and meta-analysis to plan experiments, analyze data and interpret results

**Analytical:**

It concerns the design and development of algorithms.

* ***TfidfVectorizer*** - Convert all the alphabets to vectors to make sure the model can understand the language
* Testing of Identified Approaches (Algorithms)

Various evaluation metrics can be used for this classification type of model. Some of the popular algorithms are the following:

* Logistic Regression
* k-Nearest Neighbors
* Decision Trees
* Support Vector Machine
* Random Forest Classification
* Run and Evaluate selected models

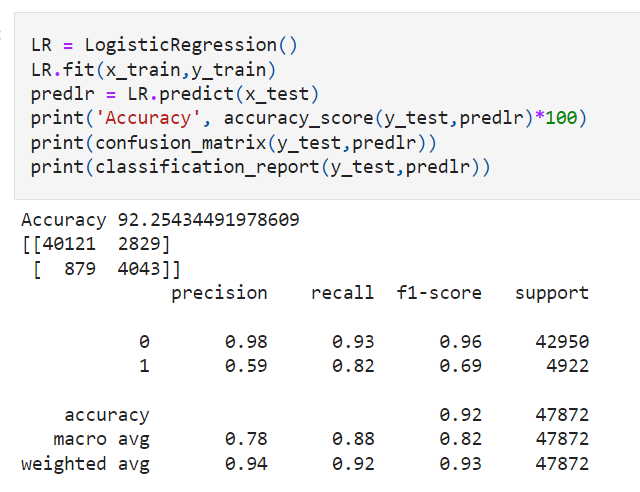
Accuracy - Percentage of correct predictions for dataset

Recall - Percentage of positive cases

Precision - Percentage of predictions are correct

F1 Score - Percentage of positive predictions are correct

***Logistic Regression:***

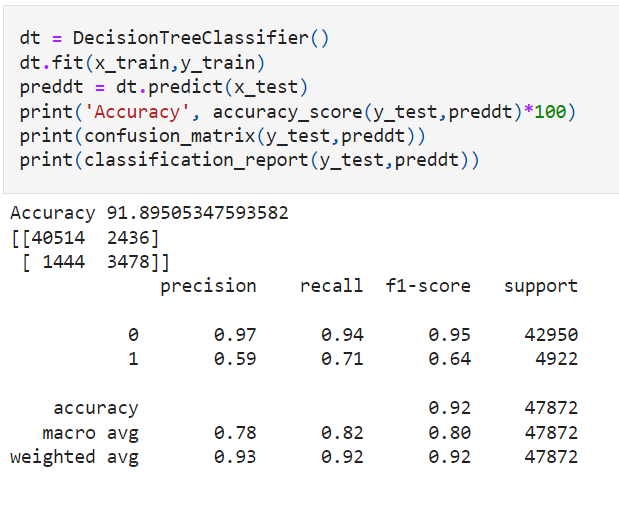


Accuracy - 88.5

Precision - Not paid - 0.61, Paid - 0.89

Recall - Not paid - 0.15, Paid - 0.99

F1 Score - Not paid - 0.25, Paid - 0.94

***Decision Tree:***

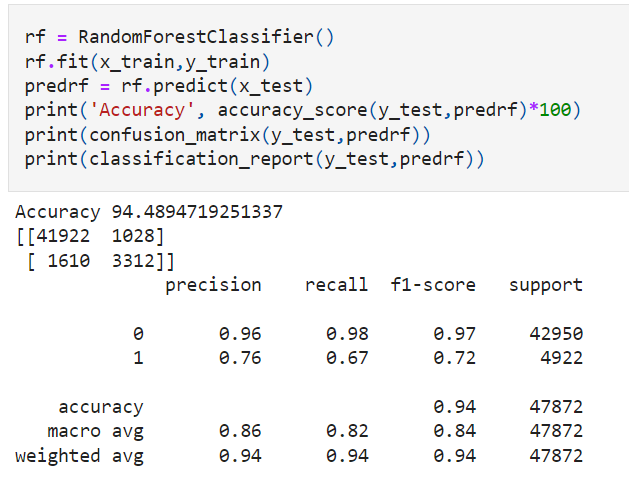
Accuracy - 85.8

Precision - Not paid - 0.42, Paid - 0.92

Recall - Not paid - 0.46, Paid - 0.91

F1 Score - Not paid - 0.44, Paid - 0.92

***Random Forest Classifier:***



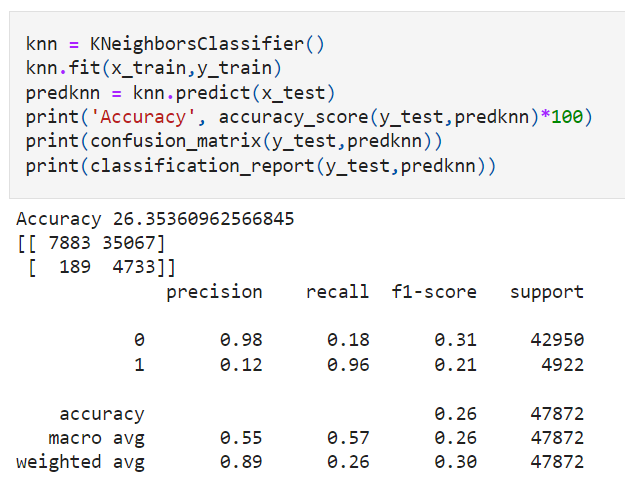
Accuracy - 90.9

Precision - Not paid - 0.72, Paid - 0.92

Recall - Not paid - 0.42, Paid - 0.98

F1 Score - Not paid - 0.53, Paid - 0.95

***KNeighbors Classifier:***



Accuracy - 90.1

Precision - Not paid - 0.64, Paid - 0.92

Recall - Not paid - 0.43, Paid - 0.97

F1 Score - Not paid - 0.52, Paid - 0.95

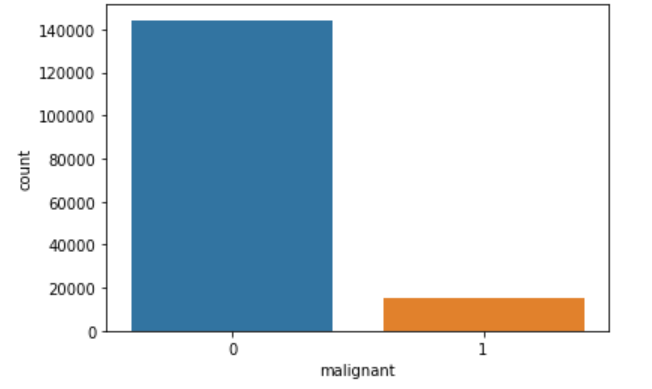
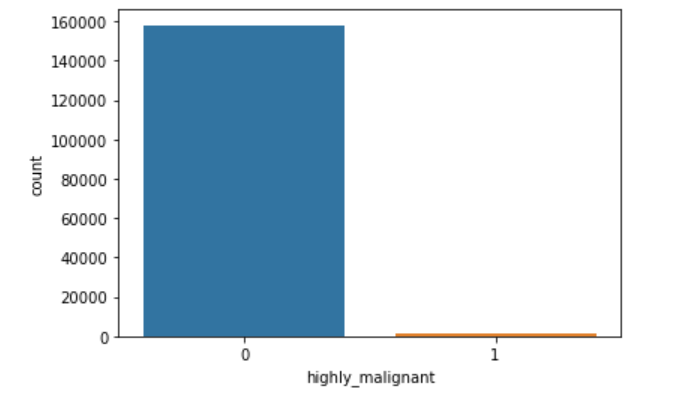
* Key Metrics for success in solving problem under consideration

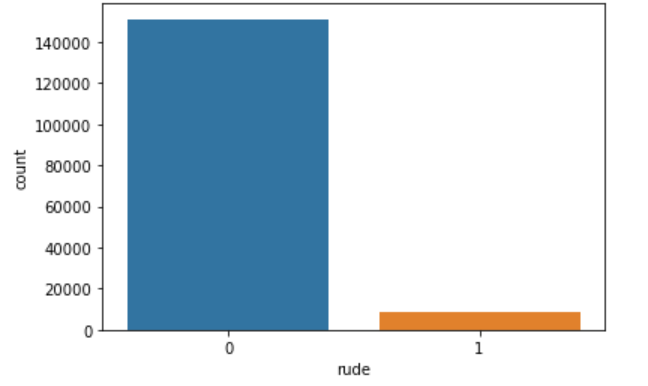
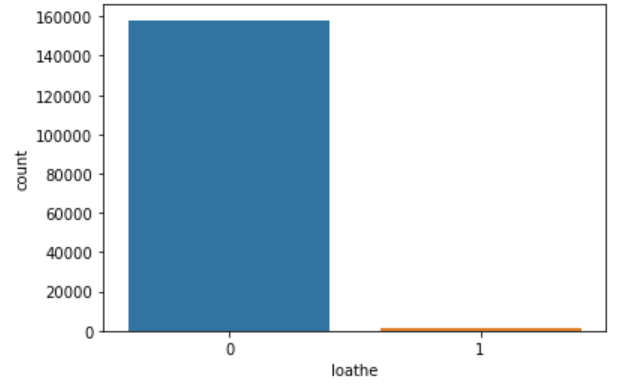


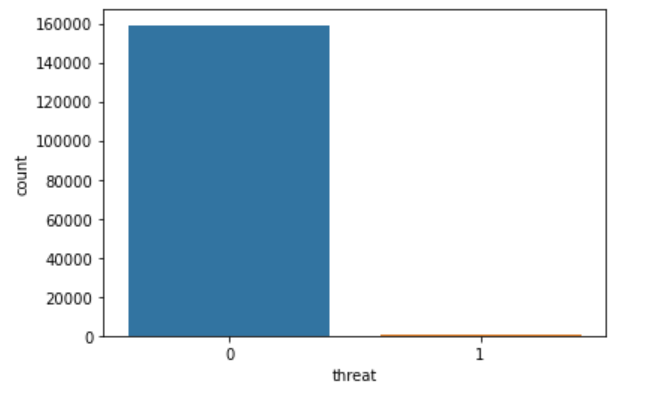
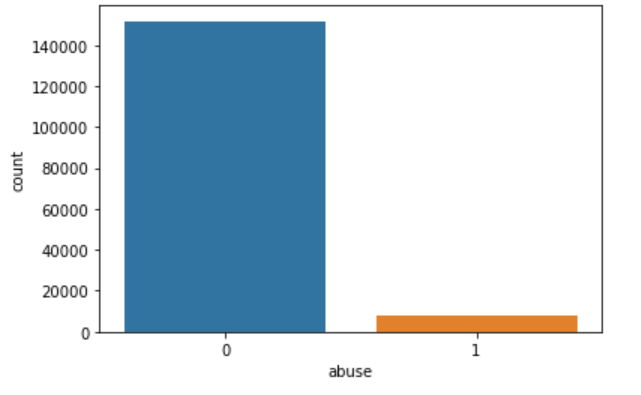
|  |  |
| --- | --- |
| **Libraries** | **Usage** |
| Pandas and numpy | Importing dataset, data cleaning, data wrangling and exploratory data analysis |
| Matplotlib and seaborn | Visualisation libraries |
| WordNetLemmatizer | To convert the similar to a unique word having the same meaning in natural language |
| stopwords | Used to check the words which does not help in determining the solution for the model |
| string | This package is used to remove the punctuations, numbers, etc |
| Accuracy score, confusion matrix, classification report | For concluding the results |
| Train\_test\_split | To separate the training and testing dataset |
| Logistic Regression, DecisionTreeClassifier, RandomForestClassifier, SVC, KNeighborsClassifier | All these are machine learning algorithms to find the results |
| Cross\_val\_score | To check the best fitting of the model |
| GridSearchCV | For hyper parameter tuning |
| TfidfVectorizer | To convert the natural language to vectors only which can be read and analysed by the models |
| Plot\_roc\_curve | To check whether the model is good by checking the area under the curve |

* Visualizations

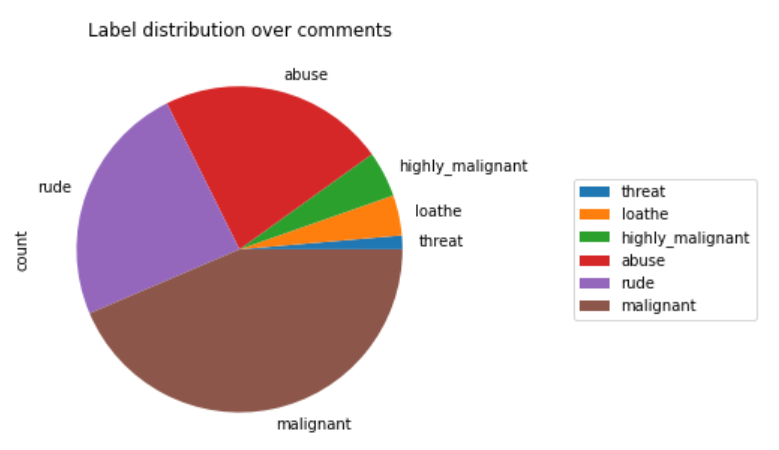
Matplotlib and seaborn is used for visualisations





Visualised the features separately for whether the comments are malignant or not and further.



The pie chart shows how much comments are malignant, threat, abusive, loathe, highly\_malignant and rude.

* Interpretation of the Results
* The features are visualized separately to understand how the data is balanced or not. Pie chart shows how much comments are to be considered for building the model because we are building the malignant comments classifier.
* In data preprocessing, duplicates are removed which helps to get unbiased result
* The words are lemmatised instead of stemming which will help us to understand the meaning of the word
* Stopwords are removed, so that the model can get a good unbiased result

**CONCLUSION**

* Key Findings and Conclusions of the Study

The relation between the input and output variables cannot be found her because we are doing a natural language processing model using natural language toolkit in python. Correlation is checked and nothing is found. The features are consolidated to find whether the comments are malignant or not to get better results. From the models, random forest classifier gives better result and also on comparing the cross validation the model fitted well.

On observing the ROC curve, the area under the curve is more and that shows that the model is good. Save the model and that can be used for predictions later since the model is trained well.

We have also got test data to predict for which we have predicted whether the comments are malignant or not.

* Learning Outcomes of the Study in respect of Data Science

Visualising the features tells the balancing of the dataset. After checking the correlation, nothing is found. The data already does not have null value and the data is pre-processed before building the model since we are building a natural language processing model.

Logistic Regression and random forest classifier is good to go with the model. Random forest classification, SVC takes time for training. SVC takes very long time to train the model. Cross-validation took time to complete and checked whether the model is underfitting and overfitting the model.

Random Forest Classifier gave 94% accuracy approximately. Even after hyper parameter tuning, the model score did not increase. So, we can use Random Forest Classifier for further predictions.

* Limitations of this work and Scope for Future Work

Data is clearly explained. Training data and testing data is given by the client. The model can be tested by using stemming during pre-processing.