

**Project Report On**

**“Malignant Comment Classification”**



**Submitted by:**

**ASHISH KUMAR RATHOR**

**ACKNOWLEDGMENT**

I would like to express my sincere gratitude for presenting this report on the “Malignant Comments Classification” project. Working on this project was a good experience that has given me basic knowledge about Machine Learning Models. This project also helped me in doing lots of research wherein I came to know about so many new things.

At the commencement of this project report, I would like to evince my deepest sense of gratitude to SME MR. MOHD.KHASIF. Without her guidance, insightful decisions, valuable comments, and corrections it would not have been possible to reach this mark.

I would like to express my gratitude to Flip Robo Technologies and Data Trained for providing a suitable environment and guidance to complete my work.

**TABLE OF CONTENTS**

**INTRODUCTION**

* Business Problem Framing
* Conceptual Background of the Domain Problem
* Review of Literature
* Motivation for the Problem Undertaken

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem
* Data Sources and their formats
* Data Pre-processing Done
* Data Inputs- Logic- Output Relationships
* State the set of assumptions (if any) related to the problem under consideration
* Hardware and Software Requirements and Tools Used

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* Testing of Identified Approaches (Algorithms)
* Run and Evaluate selected models
* Key Metrics for success in solving the problem under consideration
* Visualizations
* Interpretation of the Results

**CONCLUSION**

* Key Findings and Conclusions of the Study
* Learning Outcomes of the Study in respect of Data Science
* Limitations of this work and Scope for Future Work

**INTRODUCTION**

Over the years, social media and social networking use have been increasing exponentially due to an upsurge in the use of the internet. A Flood of information arises from m online conversations daily as people can discuss, express themselves, and air their opinion via these platforms.

Every day, we get a tremendous amount of short content data from the blast of online correspondence, web-based business, and the utilization of advanced gadgets. This volume of data requires text mining apparatuses to carry out the various report tasks in an opportune and suitable way. Detecting and controlling verbal abuse in an automated fashion is inherently an NLP task (Natural Language Processing). Text Classification is a great point for NLP.

Nowadays, every social media site and application uses a machine learning approach. Machine Learning has simplified the task that may take a long duration to complete without it. Most of the approaches require text analysis and classification techniques. Classification of the comments is necessary before posting on online platforms. This paper discusses different methodologies like logistic regression, support vector machine, multinomial naïve Bayes, etc. for comment classification into 6 different categories viz. malignant, highly malignant, rude, threat, abuse and loathe.

**Business Problem Framing**

Social media has given a lot of people beyond imagination. In this era of technology, it has become the hub of information. The numbers of content on social media are vast and rich and everything has found a place on social media that may be anything. It has given wings to its users to fly high and express their feelings. It has become a boon for mankind but we all know that if it is good there must be bad. Likewise, social media has also got a dark side.

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. 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 behavior.

There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influences are facing backlash 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 offensive.

**Conceptual Background of the Domain Problem**

In the past few years, it is seen that the cases related to social media hatred have increased exponentially. Social media is turning into a dark venomous pit for people nowadays. Online hate is the result of differences in opinion, race, religion, occupation, nationality, etc. In social media, the people spreading or involved in such kind of activities uses filthy language, aggression, images, etc. to offend and gravely hurt the person on the other side. This is one of the major concerns now.

The result of such activities can be dangerous. It gives mental trauma to the victims making their lives miserable. People who are not well aware of mental health online hate or cyberbullying become life-threatening for them. Such cases are also on the rise. It is also taking its toll on religions. Every day we can see an incident of fighting between people of different communities or religions due to offensive social media posts.

Online hate, described as abusive language, aggression, cyberbullying, hatefulness, insults, personal attacks, provocation, racism, sexism, threats, or toxicity has been identified as a major threat on online social media platforms. These kinds of activities must be checked for a better future.

**Review of Literature**

Aggression by text is a complex phenomenon, and different knowledge fields try to study and tackle this problem. In this study, several related literatures is used to express different types of aggression. Some of those are hate, cyberbullying, abusive language, malignant, flaming, threatening, extremism, radicalization, and hate speech. This research found a few dedicated works that address the effect of incorporating different text transformations on the model accuracy for sentiment classification. In this work, we performed a systematic review of the state-of-the-art malignant comment classification using machine learning methods with NLP text processing. In our analysis of every primary study, we investigated the data set used, evaluation metric, used machine learning methods, classes of malignant and non-malignant, and comment language.

**Motivation for the Problem Undertaken**

The main objective of this study is to investigate which method from a chosen set of machine learning techniques performs the best. So far, we have a range of publicly available models served through the Perspective API, including toxicity/malignant comments. But the current models still make errors, and they don’t allow users to select which type of toxicity they are interested in finding.

The project which is given by Flip ROBO as a part of the internship program gives the insight to identify major factors that lead to cyberbullying and online abusive comments. The exposure to real-world data and the opportunity to deploy my skillset in solving a real-time problem has been the primary objective. However, the motivation for taking on this project was that it is relatively a new field of research. Here we have many options but less concrete solutions. The main motivation was to classify the news to bring awareness and reduce unwanted chaos and make a good model which will help us to know such kinds of miscreants. Our goal is to build a prototype of online hate and abuse comment classifier which can be used to classify hate and offensive comments so that they can be controlled and restricted from spreading hatred and cyberbullying

**2. ANALYTICAL PROBLEM FRAMING**

**Mathematical/ Analytical Modelling of the Problem:**

We are provided with two different datasets. One for training and another one to test the efficiency of the model created using the training dataset. The training data provided here has both dependent and independent variables. As it is a multiclass problem it has 6 independent/target variables. Here the target variables are named “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.

It is a binary classification problem as the target columns give binary outputs and all independent variables have the text so it is clear that it is a supervised machine learning problem where we can use the techniques of NLP and classification-based algorithms of Machine Learning. Here we will use NLP techniques like word tokenization, lemmatization, stemming, and TF-IDF vectorizer then those processed data will be used to create the best model using various classification-based supervised ML algorithms like Logistic Regression, Multinomial NB, Random forest Classifier, Gradient Boosting Classifier, SVC, Decision Tree Classifier, and Adaboost Classifier.

**Data Sources and their formats**

The data set provided by Flip Robo was in the format of CSV (Comma Separated Values). The data set contains the training set, which has approximately 159571 samples, and the test set which contains nearly 153164 samples. All the data samples contain 8 fields which include ‘Id’, ‘Comments’, ‘Malignant’, ‘Highly malignant’, ‘Rude’, ‘Threat’, ‘Abuse’, and ‘Loathe’. In the particular dataset, all the columns are of object data type. The attribution information is as follows:

The data set includes:

* **Malignant:** It is the Label columns 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 an 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 hat comments 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 Done**

Data pre-processing is the process of converting raw data into a well-readable format to be used by the Machine Learning model. Data pre-processing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, we must pre-process our data before feeding it into our model. I have used the following pre-processing steps:

* Importing necessary libraries and loading the dataset as a data frame.
* Checked some statistical information like shape, the number of unique values present, info, null values, value counts, duplicated values, etc.
* Checked for null values and did not find any null values. And removed Id.
* Done feature engineering and created new columns viz label: which contain both good and bad comments which is the sum of all the labels, comment length: which contains the length of comment text.
* Visualized each feature using seaborn and matplotlib libraries by plotting, counting plot, and word cloud for each label.
* Checked correlation using heatmap.
* After getting cleaned data used the TF-IDF vectorizer. It’ll help to transform the text data into a feature vector that can be used as input in our

modeling. It is a common algorithm to transform the text into numbers. It measures the originality of a word by comparing the frequency of appearance of a word in a document with the number of documents the words appear in.

Mathematically,

**TF-IDF = TF(t\*d)\*IDF(t,d)**

**Hardware & Software Requirements & Tools Used**

To build machine learning projects it is important to have the following hardware and software requirements and tools.

|  |  |
| --- | --- |
| **Hardware** | Processor: core i5  RAM: 12 GB  ROM/SSD: 512 GB |
| **Software** | Distribution: Anaconda Navigator  Programming language: Python  Browser-based language shell: Jupyter Notebook |

**import seaborn as sns: Seaborn** is also a Python library used for plotting graphs with the help of Matplotlib, Pandas, and Numpy. It is built on the roof of Matplotlib and is considered a superset of the Matplotlib library. It helps in visualizing univariate and bivariate data.

With the above sufficient libraries, we can perform pre-processing and data cleaning, and model building.

**3. MODEL/S DEVELOPMENT AND EVALUATION**

**Identification of possible Problem-solving approaches (Methods):**

In this project 6 features define the type of comment like malignant, hate, abuse, threat, and loathe but we created another feature named “label” which is combined all the above features and contains the labeled data in the format of 0 and 1 where 0 represents “NO” and 1 represents “Yes”. In this NLP-based project, we need to predict the multiple labels which are binary. I have converted text into feature vectors using TF-IDF vectorizer and separated our features and labels. Also, before building the model, I made sure that the input data is cleaned and scaled before it was fed into the machine-learning models.

**Testing of Identified Approaches (Algorithms)**

Since the target variable is categorical in nature, from this I can conclude that it is a classification-type problem hence I have used the following classification algorithms. After the pre-processing and data cleaning, I left with 10 columns including targets. The algorithms used for training the data are as follows:

1. Logistic Regression
2. MultinomialNB
3. Random forest Classifier
4. SVC
5. Gradient Boosting Classifier
6. Decision Tree Classifier
7. Gradient Boosting Classifier (XGB)
8. AdaBoost Classifier
9. KNeighbors Classifier

**Model building:** After cleaning and processing data, I performed a train test split to build the model. I have built multiple classification models to get the accurate accuracy score, and evaluation metrics like precision, recall, confusion matrix, f1 score, log loss, and hamming loss. I got the Random Forest classifier as the best model which gives a 95.66% accuracy score. Finally, I saved my final model and got good prediction results for the test dataset.

# CONCLUSION

## Key Findings and Conclusions of the Study

From the above analysis the below-mentioned results were achieved which depict the chances and conditions of a comment being a hateful comment or a normal comment;

✓ With the increasing popularity of social media more and more people consume feeds from social media and due to differences, they spread hate comments instead of love and harmony. It has strong negative impacts on individual users and broader society.

The conclusion for our study:

* In the training dataset, we have only 10% of is spreading hate on social media.
* In this 10% data most of the comments are malignant, rude, or abusive.
* After using the word cloud, we find that there are so many abusive words present in the negative comments. While in positive comments there is no use of such comments. o Some of the comments are very long while some are very short

## Learning Outcomes of the Study in respect of Data Science

While working on this project we learned many things and gains new techniques and ways to deal with uncleaned text data. Found how to deal with multiple target features. Tools used for visualizations give a better understanding of the dataset. We have used a lot of algorithms and found that in the classification problem where we have only two labels, XGB Classifier gives better results compared to others.

It is possible to classify the content of the comments into the required categories of authentic, however, using this kind of project awareness can be created to know what is fake and authentic.

## Limitations of this work and scope for future work

**Limitations:** This project was amazing to work on, it creates new ideas to think about but there were some limitations in this project like an unbalanced dataset. Every effort has been put into it for perfection but nothing is perfect and this project is of no exception. Certain areas can be enhanced**. Future work:** In future work, we can focus on performance and error analysis of the model as lots of comments are misclassified into the hate category. Previous work has achieved success using various algorithms on data in the English language but in the future, we can consider having data in regional languages. We can also work on after work of the detection of malignant comments like automatic blocking of the user, and auto-deletion of harmful comments on social media platforms. Comment detection is an emerging research area with few public datasets. So, a lot of work needs to be done in this field