MALIGNANT COMMENT CLASSIFICATION USING MACHINE LEARNING MODEL

BY

Name : Fokhrul Islam
ID: 191-15-12325
Name : Chanchal Ray
ID : 191-15-12708

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project titled **"Malignant Comment Classification using machine learning model."**, submitted by Fokhrul Islam, ID No: 191-15-12325 and Chanchal Ray, ID No: 191-15-12708 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 25/01/2023.

BOARD OF EXAMINERS

____

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman

____

Dr. Md. Monzur Morshed
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

____

Internal Examiner

Dewan Mamun Raza
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

25.1.2023

____

Dr. Ahmed Wasif Reza
Associate Professor
Department of Computer Science and Engineering
East West University

External Examiner

DECLARATION

I hereby declare that, this project has been done by me under the supervision of **Md Zahid Hasan, Associate professor, Department of CSE,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

____

Dr. Md Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

____

Md. Sadekur Rahman,
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Fokhrul
____
Fokhrul Islam
ID: 191-15-12325
Department of CSE
Daffodil International University

____
Chanchal Ray
ID: 191-15-12708
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to Almighty Allah for His divine blessing which makes me possible to complete the final year project successfully.

I really grateful and wish my profound indebtedness to **Md Zahid Hasan**, **Associate professor**, Department of CSE, Daffodil International University, Dhaka, deep knowledge & keen interest of my supervisor in the field of Machine Learning to carry out this project. His endless patience, scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Touhid Bhuiyan**, Head**,** Department of CSE, for his kind help to finish our project and also to other faculty members and the staffs of CSE department of Daffodil International University.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Online comments that are visible in public spaces typically contain a big percentage of constructive comments, but a sizeable percentage also contain toxic comments. Online datasets are collected and cleaned of noise. As a result of the large number of errors in the comments, which greatly increases the number of features, before feeding the dataset to the classification models utilizing the term frequency-inverse document frequency (TF-IDF) approach, the machine learning model must first turn it into transformed raw comments for training.Six different machine learning techniques use for classify the dataset.The logistic regression algorithm is used to train the processed dataset. Decision tree classifiers use for visualize data.Random forest classification ,XGB Boost,AdaBoost Classifier,and KNN this model gives best accuracy.Then using confusion metrics for their prediction.We have applied six different machine learning techniques, such as logistic regression, decision trees, random forest classification, XGB Boost, AdaBoost Classifier, and KNN, to our dataset and got the accuracy of 0.95, 0.99, 0.99, 0.96, 0.95, and 0.92, respectively. Random forest classification and decision tree classifiers got an accuracy of 0.99, which was the highest among all classifiers.

TABLE OF CONTENTS

CONTENTS **PAGE**

Board of examiners i

Declaration ii

Acknowledgements iii

Abstract iv

List of Figures viii

List of Tables ix

CHAPTER 1: Introduction **1-4**

1.1 Introduction 1

1.2 Motivation 1

1.3 Rational of the Study 2

1.4 Expected Outcome 2

1.5 Research Question 3

1.6 Report Layout 3

CHAPTER 2: Background **5-7**

2.1 Preliminaries/Terminologies 5

2.2 Related Works 5

2.3 Comparative Analysis and Summary 6

2.4 Scope of the Problem 7

2.5 Challenges 7

CHAPTER 3: Research Methodology **8-17**

3.1 Research Subject and Instrumentation 8

3.2 Data Collection Procedure/Dataset Utilization 11

3.3 Statistical Analysis 11

3.4 Vectorization 12

3.5 Proposed Methodologies/Applied Mechanism 13

3.6 Confusion matrix 17

3.7 Implementation Requirements 17

CHAPTER 4: Experimental Result and Discussion **18-23**

4.1 Experimental Setup 18

4.2 Experimental Result 18

4.3 Experimental Analysis 19

4.4 Discussion 22

CHAPTER 5: Impact on Society, Environment and Sustainability **24-25**

5.1 Impact on Society 24

5.2 Impact on Environment 24

5.3 Ethical aspects 24

5.4 Sustainability Plan 25

CHAPTER 6: Summary, Conclusion and Implication For Future Research **26-27**

6.1 Summary of the Study 26

6.2 Conclusions 26

6.3 Implication for Further Study 26

REFERENCES **28-29**

LIST OF FIGURES

FIGURES	PAGE NO
Figure 1: Proposed methodology	9
Figure 2 : Comment classes	12
Figure 3 :The working of Random Forest Classification	14
Figure 4 : Structure of XGBoost	15
Figure 5 : AdaBoost Classifier Working Process.	16
Figure 6 : Confusion matrix of (a) Logistic Regression (b) Decision Tree Classification (c) XGBoost classification (d) AdaBoost Classifier (e) KNN Classification (f) Random Forest Classification	21
Figure 7 : Plotting Graph of (a) Logistic Regression (b) Decision Tree Classification (c) XGBoost classification (d) AdaBoost Classifier (e) KNN Classification (f) Random Forest Classification	22
Figure 8 : Features or words that make a comment toxic	23

LIST OF TABLES

TABLES	PAGE NO
Table 1: Comparative summary and analysis	7
Table 2: Datasets Description	10
Table 3: Model Performance (Supervised Machine Learning)	19

CHAPTER 1

INTRODUCTION

1.1 Introduction

A significant issue is the presence of toxic and irrelevant remarks on social media and other online platforms that connect people all over the world. Some people are prone to losing control over any given topic, at which point they may post anything offensive or racially charged, which may leave the recipient with a sense of being harassed or abused. These are the types of things that are typical on social media; people often use it to voice their opinions and protests, and some platforms even allow for real-time communication between users. Even though there are some people who aren't very toxic, unfortunately there are also going to be some poisonous people in this world.

Therefore, the purpose of this project is to identify those harmful or toxic comments and posts in order to train artificial intelligence to automatically detect them for us. By using machine learning and algorithm data analysis, we are able to determine how harmful and malicious a comment may be, thereby allowing us to identify the individuals who made it.

Also his project's objective is to make the internet a more secure place to conduct business. When harmful comments are located on social media platforms, users have the ability to quickly and simply complain and have them removed. In this increasingly digitized environment, this would make it possible for people to have stronger connections with one another over the long term.

1.2 Motivation

Using this analysis of the data, we attempted to determine the level of danger posed by a comment as well as the level of danger posed by that comment. By these means, we attempted to build it as an AI base by utilizing the facts in a way that would result in a conclusion that is just regarding the comment or the post.

We sought to make it a secure and better location to avoid the poisonous environment that can sometimes be found in social media, which may lead to us getting involved in some major controversy in a foreign country. This is because AI and ML are being used by many various things.

It is possible for one's job stability and future work chances to be negatively impacted by the disinhibiting effect of using the internet. Indicatively, the research conducted by the Wikimedia foundation indicated that 54 percent of people who had been subjected to online abuse reported a reduction in their engagement in the specific project that had taken place [1]. Therefore, this work may be beneficial for individuals who do not feel safe or who have been harassed on social media or any other platform that connects people with words.

1.3 Rationale of the Study

There have been a significant number of research papers published in recent years on the problem of classifying toxic comments; however, to this day, there has not been a systematic literature review of this research theme. Because of this, it is difficult to evaluate the level of development, the trends, and the research gaps. In this particular piece of study, our primary objective was to uncover prospective avenues for future research by methodically listing, contrasting, and classifying the previous research that has been conducted on the classification of poisonous comments. The findings of this comprehensive evaluation of the relevant literature will be helpful to researchers as well as practitioners of natural language processing.

1.4 Expected Outcome

We hope that the results of this data analysis will lead us to a secure roming on the internet. The social media and a better culture both deserve a more positive environment in which there is neither toxicity nor hostility.

We have applied some machine learning to comprehend the fundamental type of threads that are dangerous and evil or malicious based on the data that we have obtained. The malicious word will be detected by the AI on its own, and either a sensor will be applied to it or it will be removed.

Although there are some comments that are malicious, the overall tone of the conversation is not particularly racist or aggressive. Because of this, our system has eight fields that the AI can analyze and divide the comments into seven categories.

- Comments
- Malignant
- Highly malignant
- Rude
- Threat
- Abuse
- Loathe

Therefore, by using these classes, it will determine how a word is and how that word is specified.

1.5 Research Questions

Throughout the study process, several questions about the work arise. The main questions of our research are given below:

- How to collect data?
- How to apply preprocessing techniques?
- How to execute Machine learning model?
- How to train Machine learning model?
- How to analysis experiments and result?

1.6 Report Layout

This research paper contains total 6 chapters as given below:

Chapter 1:In this chapter provides background information about the study, including its purpose, justification, methodology, research questions, and anticipated results.

Chapter 2:Contains a discussion of the background, scope, difficulties, and solutions to the topic, as well as an overview of relevant works.

Chapter 3:Includes the research process, data collection method, analysis, and feature implementation.

Chapter 4:Including numerical and graphic representations of the results of the research, as well as experimental evaluation and some pertinent discussions.

Chapter 5:Examines the societal effects of this research.

Chapter 6: Discussing a brief overview of this study's findings, as well as a discussion of its limitations and suggestions for further research.

CHAPTER 2

BACKGROUND

2.1: Preliminaries/Terminologies:

In this research, they apply a machine learning method to natural language processing to classify and detect poisonous language in online user comments. Using these parameters, we obtain a Mean Validation Accuracy of 98.08%, which is the highest numerical accuracy achieved by any Comment Toxicity Detection Model to date. This paper's research was undertaken to encourage open and honest discussion and debate in social media. If the Machine Learning Algorithms for each pipeline are utilized to deliver more accurate classifications and better results, then the Grid Search Algorithm used to the same dataset may produce a more robust model.[11]

With a hamming loss of 3.6 compared to SVM's 4.36, a paper based on the results concludes that the Binary Relevance approach with Multinomial Naive Bayes is an efficient algorithm that meets our purpose. The hammering loss for this strategy is 3.6.[3]

In terms of accuracy across multiple labels, LSTM performed best. To put it simply, it was superior to the alternatives. They predicted that the SVM and logistic regression implementations would achieve similar levels of accuracy, but they were taken aback when XGB boost did not outperform them. They pondered whether, in the long run, word embedding or character embedding would prove more reliable. They also enjoy comparing a CNN model built with LSTM code. LSTM models were also not used by them. They would experiment with LSTM models if I had adequate memory to do so. From the get-go, they opted to develop a bidirectional LSTM and tweaked the model's hyperparameters for maximum performance. Different models couldn't be tried.[4]

2.2 Related work

There have been a vanishingly small number of attempts made to classify malignant comments. In this piece of research, we present a method for developing a forecast and identifying malicious comments based on datasets.

Jigsaw and Google's Conversation AI team has been developing strategies and technologies to foster productive dialogue.[2]

LSTM was the top multi-label classifier in the dataset.It beat the competition. I expected the SVM and logistic regression implementations to have comparable levels of accuracy, but I expected XGB boost to do better and was shocked that it did not.

This paper arthur was curious if character embedding is more accurate than word embedding in the future. I'd also like to test a CNN model against the LSTM implementation. I didn't implement LSTM models either. If given enough RAM, I'd play with LSTM models. I chose to create a bidirectional LSTM from the start and fine-tuned the model hyperparameters. I couldn't try different models.[10]

For their 2017 refining model, Yu and Wang [5] suggested moving word vectors closer to emotionally comparable words and farther from emotionally dissimilar words. Using the proposed strategy, they were able to show that improvements above baseline word embeddings were possible through experimental work.

Methods for detecting damaging comments using deep learning have been researched extensively. The ensemble created by Van Aken et al.[6] outperformed every single model they tested on a whole new, massive training dataset. The best results were seen with CNN and LSTM models. Different neural network techniques for comment classification have been intensively researched in recent literature[15,16].

2.3 Comparative Analysis and Summary

In this paper, we showed a comparison study that was based on research that had already been done in the same area. Where we talked about some important things, like how accurate their study was and which algorithm they used if they used more than one. We also talked about classifiers, the method they used, and the language of the dataset etc.

Table 1: Comparative summary and analysis

Author name	Method	Classifiers	Accuracy
Kevin Khieu Neha	Binary Classification	CNN	0.889%
P. A. Ozoh, A. A. Adigun, M. O. Olayiwola	Multi-label Classification	LR	99.21%
Jakaria Rabbi	Classification	Naive Bayes	80.57%
Prinslou Tare	Classification	LSTM	0.97%

2.4 Scope of the Problem

- ❖ Applying machine learning techniques to identify the Accuracy Scores.
- ❖ If you do not have all the information, then the accuracy score does not meet your satisfaction.
- ❖ Data preprocessing
- ❖ Purifying malignant and normal comment.
- ❖ Having correct accuracy.

2.5 Challenges

- ❖ Choosing the right hardworking teammate.
- ❖ Choosing the appropriate topic for the research.
- ❖ Data collection.
- ❖ Choosing the right methodology.
- ❖ Choosing the right machine learning model.
- ❖ Dealing with data

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research subject and instrument

This section will illustrate our approach to classifying malignant comments. There have been a few studies that use machine learning to classify malignant comments. We are trying to use machine learning algorithms for maximum accuracy. Our model's workflow is depicted in Figure-

3.1.1 Proposed Methodology:

Our proposed methodology is shown below:

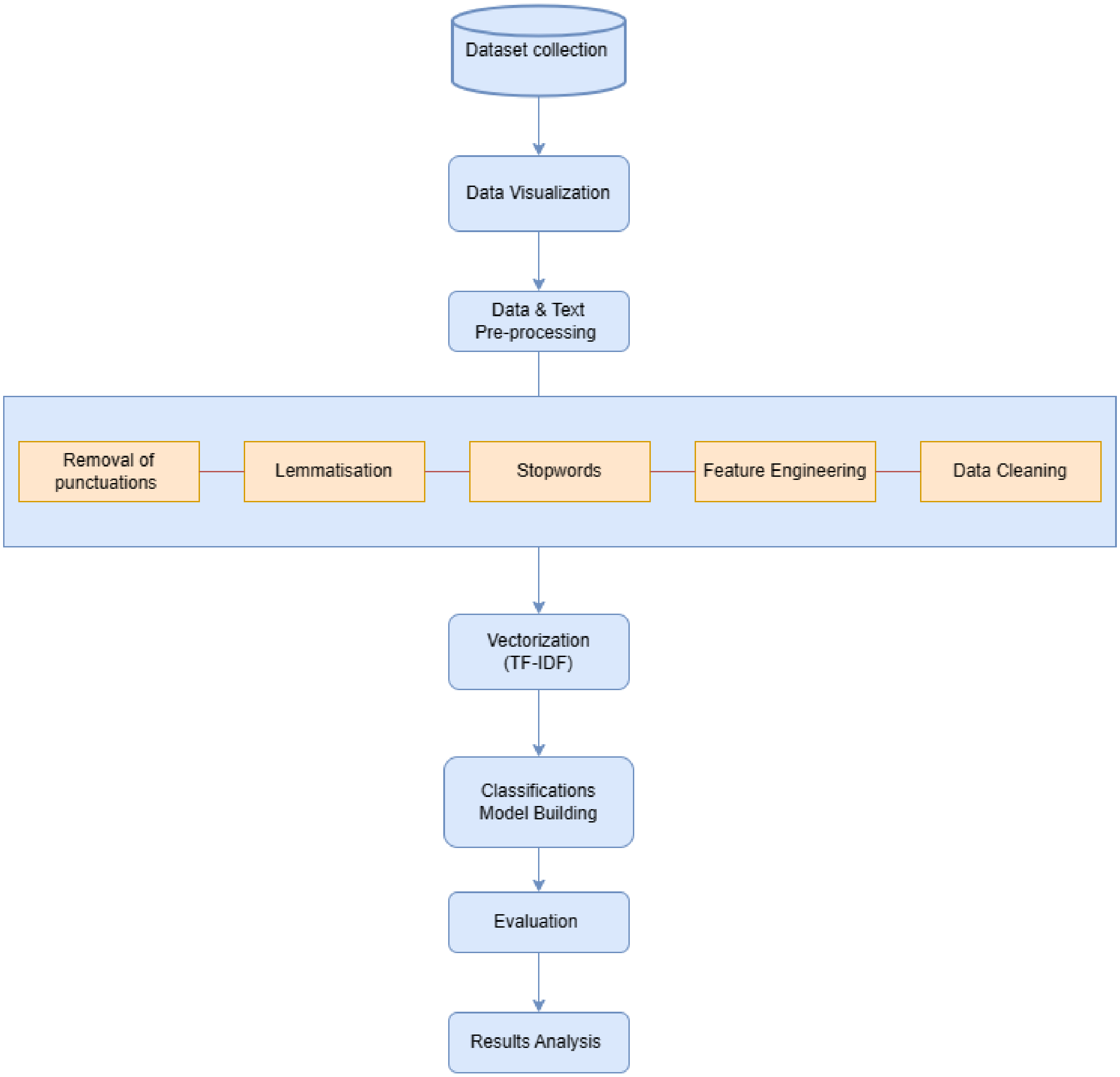


Figure 1: proposed methodology

3.1.2 Problem declaration: The social media post is designated by $I = \{w1, w2, \ldots, wn\}$, where n is the sentence length. For any input sequence I, the task is to identify the class label $ci \in C$, where $C = \{non - malignant, malignant\}$ and for each word $wi \in I$ assign a tag $y_i^S \in Y^S$, where $Y^S = \{B\text{-}T, I\text{-}T, O\}$to predict malignant span(rationale) of input sequence. To predict malignant spans, the *BIO* tagging scheme is used, where *B-T* (Begin) represents the first token in a malignant span, *I-T* (Inside) represents the inside and end tokens in a malignant span, and *O* represents the no-malignant tokens [13].

3.1.3 Data collection: We used a dataset from Kaggle [12]. The training set, which contains over 1,59,000 samples, and the test set, which contains nearly 1,53,000 samples, make up the data set. The eight fields present in all data samples are "Id," "Comments," "Malignant," "Highly Malignant," "Rude," "Threat," "Abuse," and "Loathe." There are various comments that have multiple labels. The first attribute is a unique ID associated with each comment.

Table 2: Datasets Description

Attribute	Value
Malignant	15294
Highly Malignant	1595
Rude	8449
Threat	478
Abuse	7877
Loathe	1405

3.1.4 Data preprocessing: We used the same balanced dataset that Kaggle provided and processed the data in Python.[12]

3.1.5 Text preprocessing : As a first step in this preprocessing, we stripped the comments of any punctuation and other special characters. Then we realized we needed to get rid of the worthless stop words that were included in the dataset. They have no bearing on the discussion at hand. Words were also stemmed and lemmas were created. Lemmas refer to the inflected forms of words, such as the many verb tenses, singular/plural forms, etc. Inflected variants of "gone" include go and gone, both of which are lemmas. Lemmatization refers to the process of classifying these lemmas into larger categories. That's why we lemmatize all the feedback we get.

To determine whether comments are malignant, we performed an exploratory data analysis and discovered that numerous characteristics outside the words themselves may be beneficial. Character count, proportion of capitalized letters, average word length, exclamation and question mark counts, and the total number of each were all attributes I contributed to the dataset.

Replaced email address with email, replaced phone numbers with phonenumber,Replace URLs with webaddress.We cleaned the data using regex, matching patterns in the comments and replacing them with more organized counterparts. We removed any spaces, line breaks, contractions, etc. Cleaner data leads to a more efficient model and higher accuracy.

3.2 Data Collection procedure/ Dataset Utilized : Each supervised learning algorithm requires a massive amount of data. The larger the dataset, the more accurate the result. Additionally, we require a sizable amount of data for our model. Our data was obtained from Kaggle. The dataset contains 159,000 samples.

3.3 Statistical Analysis : Exploration of Data There are 159,571 comments in this dataset. The data consists of one input feature, the string data for the comments, and the labels "Malignant," "Highly Malignant," "Rude," "Threat," "Abuse," and "Loathe" for the various kinds of malignant comments. There are all different kinds of comments in the figure on the following page. As we can see, not all comments with other labels are malignant, despite the fact that the majority of them are. It's not near enough to be a labeling error, but only "very malignant" is obviously a subclass of "malignant." This suggests that the term "malignant" is not a blanket description but rather a subcategory with significant overlap.

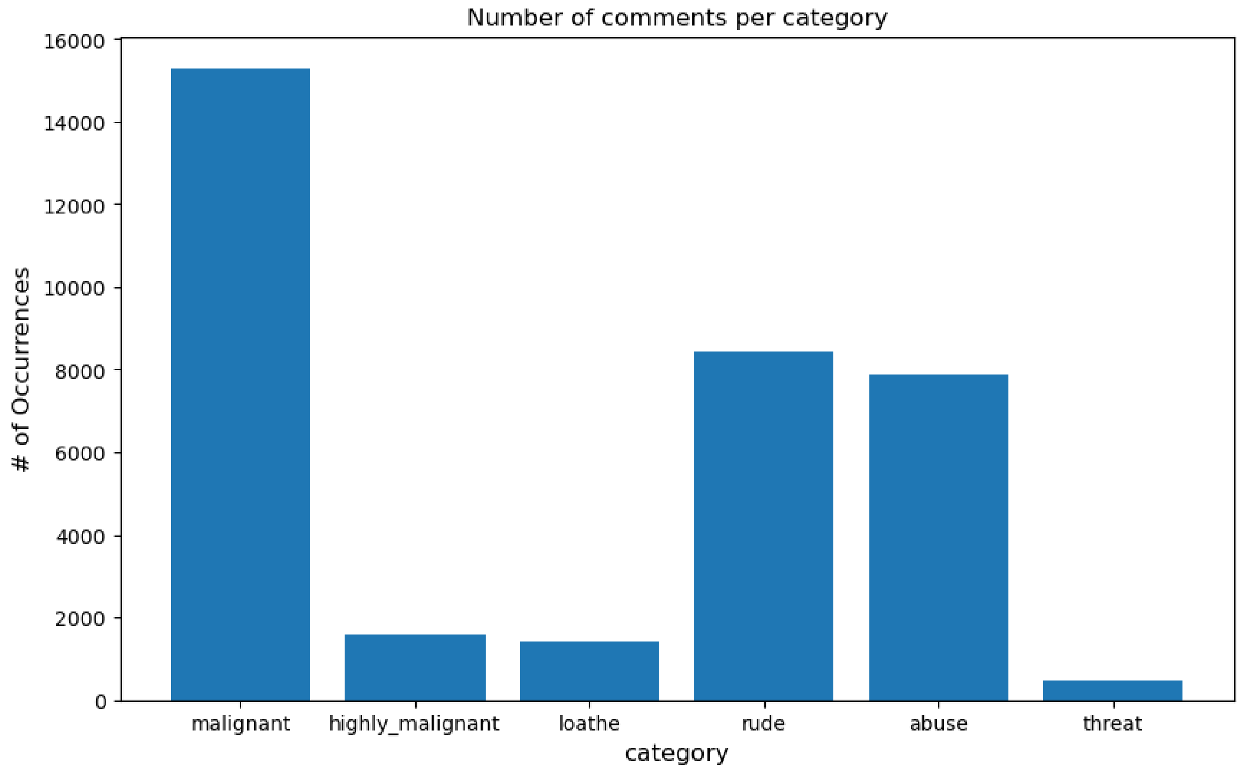


Figure 2 : Comment classes

3.4 Vectorization : We use a term frequency-inverse document frequency (tf-idf) statistic to "vectorize" the text. The existence of character n-grams and the total amount of features is a tunable parameter for model optimization.

For information retrieval issues, the TF-IDF weighted model is frequently employed. The goal is to use the frequency of words rather than the precise sequence in which they appear in the text to generate vector models. In this case, let's assume there is a dataset of N text documents. D, TF, and IDF shall be defined as follows in all relevant documents: Frequency over a Term (TF): The term frequency (TF) for a term "t" is defined as the number of occurrences of "t" in a document "D".

To calculate the Inverse Document Frequency (IDF) of a word, take the logarithm of the ratio of the total number of documents in the corpus to the number of documents that include the term T.[11]

3.5 Proposed Methodology /Applied Mechanism : After the data has been cleaned and a random train-test split has been applied, the algorithms and steps involved in the proposed malignant comment classification system are discussed. These include "logistic regression," "decision tree classifier," and "random forest classifier," all of which aid in the classification of the comments and yield a conclusive result. We use six different models that I'll discuss below.

3.5.1 Logistic Regression : For supervised training, we turn to the logistic regression (LR) technique. The first step we did was to use the logistic regression technique in our analysis. The goal was to provide a foundation from which to build. Since this is a multi-class classification issue, we used the normalized GloVe word embeddings for all terms as input to the Sklearn default logistic regression package and a one-vs-all classifier to execute logistic regression. Logistic regression makes use of the formula

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

Where x represents that features. Implementing this algorithm, I was able to obtain Training accuracy is 0.95 and Test accuracy is 0.95 .[10]

3.5.2 Decision tree classifiers visualize data : Then we used Decision Tree Algorithm for visualize data.After tf-idf transformation, a complete numeric featured dataset is obtaine.The data set is split between the training and testing part we now apply the decision tree model on the training set; predict the results on the training and testing set both and then check the accuracy.The best feature of the dataset is placed at the root of the tree.The Training Samples are splitted into subsets such that each subset contains data with the same value for a feature.The model accuracy as obtained in the training and testing data set .

Training accuracy is 0.99 and Test accuracy is 0.93.

3.5.3 Random Forest Classifier : The Random Forest (RF) classifiers are suitable for dealing with the high-dimensional noisy data in text classification. An RF model comprises a set of decision trees, each of which is trained using random subsets of features. The random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set, and then it collects the votes from different decision trees to decide the final prediction. The great thing about the Random Forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction.

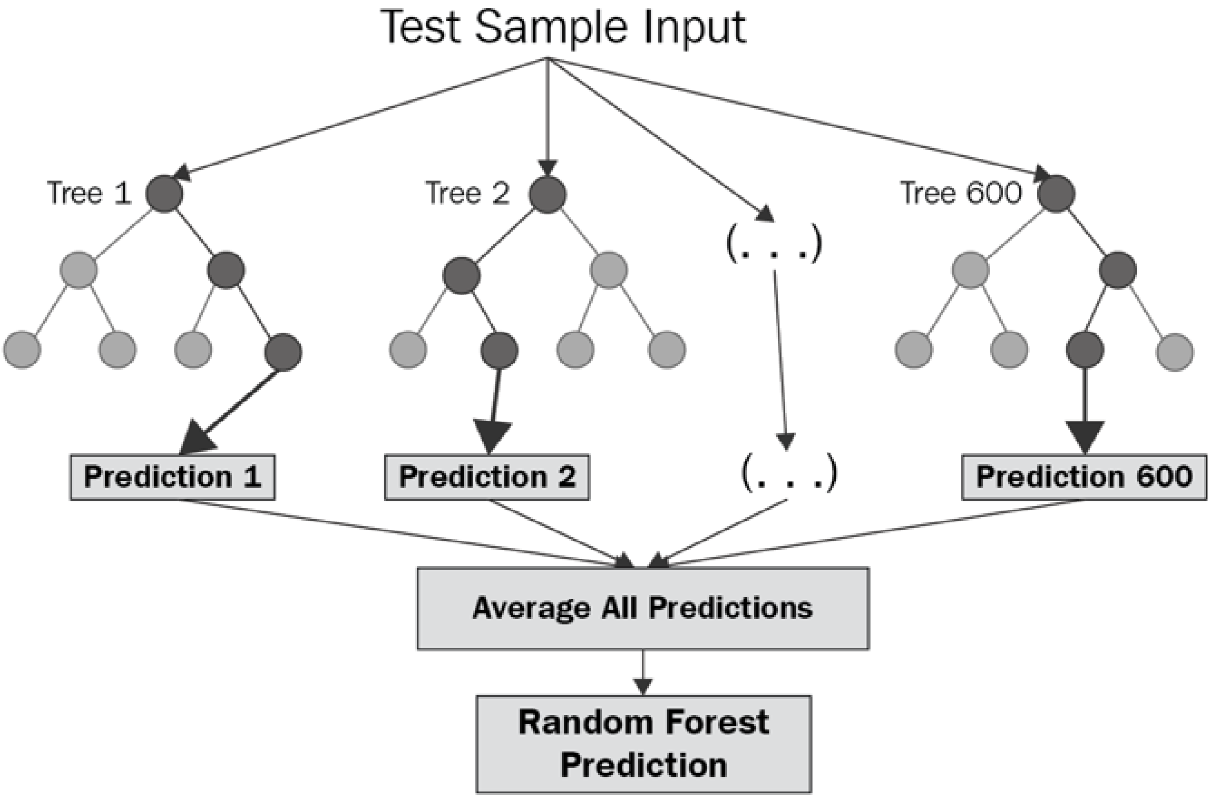


Figure 3 :The working of Random Forest Classification.

3.5.4 XGB Boost : XGBoost is a library of gradient-boosting techniques that have been tuned for distributed use[9].XGBoost is a gradient-boosted decision tree implementation that was developed for speed and performance. The method for putting this model into action was very much like the one used for the logistic regression models.XGBoost's basic structure is shown in Figure4 . We used the one-versus-all classifier that comes with the Sklearn default LR package and fed it the normalized GloVe word embeddings for each and every word as the input. We decided to take this strategy because we believed that the model's singular concentration on computational speed and overall performance would be beneficial in enhancing the predicted accuracy of our analyses. Additionally, the model is quite helpful for predictive modeling in categorization, particularly when working with structured or tabular datasets. This is one of the applications in which it excels.

Training accuracy is 0.96 and Test accuracy is 0.95.

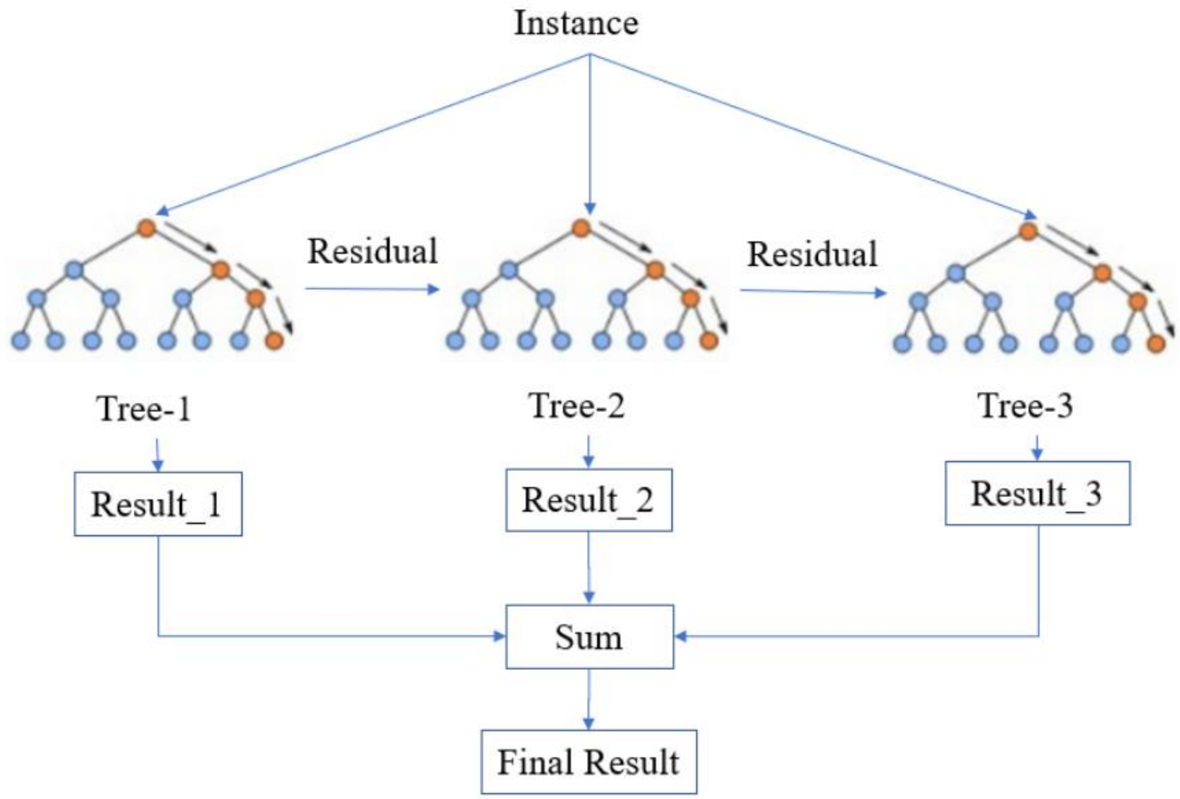


Figure 4 : Structure of XGBoost.

3.5.5 AdaBoost Classifier : To increase classifier accuracy, the AdaBoost Classifier combines many classifiers. An iterative ensemble algorithm is AdaBoost. Through the combination of several ineffective classifiers, the AdaBoost classifier creates a powerful classifier with high accuracy. The fundamental idea underlying Adaboost is to train the data sample and set the classifier weights in each iteration to provide precise predictions of uncommon occurrences. This supervision approach is rather straightforward, has strong generalizability and high classification accuracy, which can limit the overfitting of the model to some extent. The training subset used by Adaboost is shown to be chosen at random in Figure 5. It iteratively trains an AdaBoost ML model by selecting a new training set based on the accuracy of the previous training. For the next round, it provides more weight to observations that were mistakenly categorized. Additionally, in each iteration, the trained classifier is given more or less weight depending on how well it performed. For classification purposes, greater importance will be placed on the classifier that achieves higher accuracy. The training set is iterated over and over until a good fit is found, or until the maximum number of estimators is achieved. Cast your vote among the many classification algorithms you've developed.

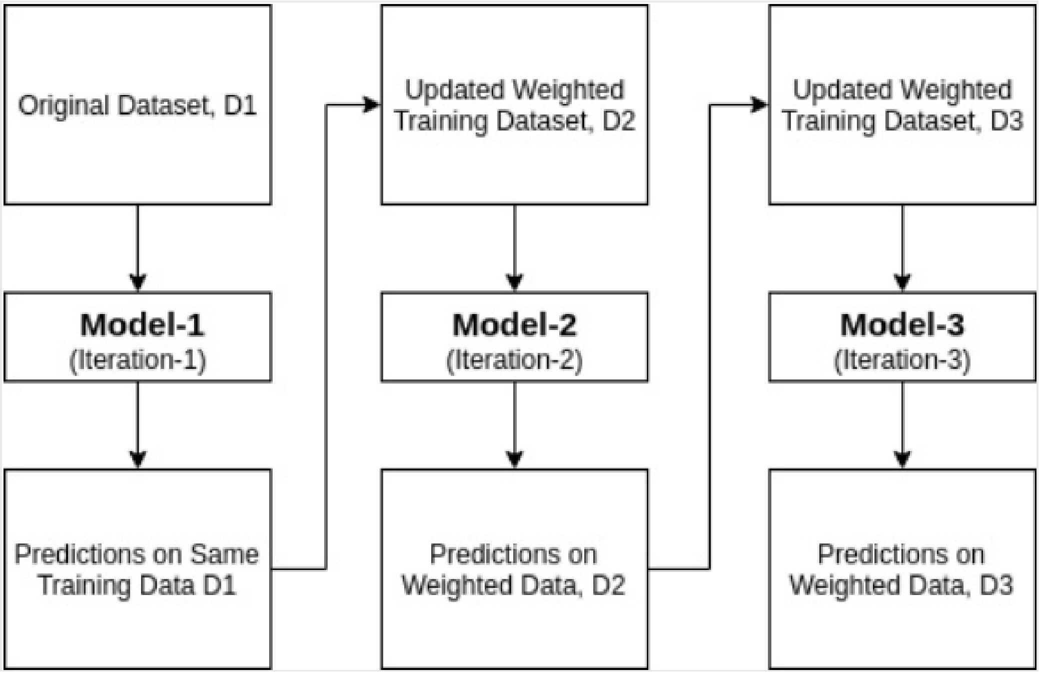

Original Dataset, D1
Updated Weighted Training Dataset, D2
Updated Weighted Training Dataset, D3
Model-1 (Iteration-1)
Model-2 (Iteration-2)
Model-3 (Iteration-3)
Predictions on Same Training Data D1
Predictions on Weighted Data, D2
Predictions on Weighted Data, D3

Figure 5 : AdaBoost Classifier Working Process.

3.5.6 KNN : K is the number of nearest neighbours in KNN. The primary determining factor is the number of neighbors. K is generally an odd number if the number of classes is 2. The algorithm is referred to as the nearest neighbor algorithm when K=1.Assume P1 is the point for which the label must provide a prediction.Finding the nearest point to P1 comes first, followed by labeling the closest point that is associated with P1.Consider P1 is the point for which the label must provide a prediction.To classify points, we first determine the k points that are closest to P1, and then we divide the votes among those k points. Each object casts a vote for the class they belong to, and the prediction belongs to the class with the most votes.[14] With KNeighborsClassifier we got : Training accuracy is 0.92 and Test accuracy is 0.91.

3.6 Confusion matrix

A method for summarizing the effectiveness of the classification algorithm is the confusion matrix. An improved understanding of the categorization model's strengths and weaknesses can be obtained by computing a confusion matrix. The following are the processes for building a confusion matrix:

- Get a dataset with predicted outcome values for testing or validation.
- For each row in the test dataset, predict the value.
- Count the number of accurate predictions for each class based on the predicted results and predictions.

It is a performance measurement for a machine learning classification problem where the output can be two or more classes. It is a table with four different combinations of predicted and actual values. True positives are predicted positives that occur. True negatives are the predicted negatives, and they're true. False Positives are defined as positives that are false. False negatives are the predicted negatives, and they're false. predicted values as "positive" and "negative," and actual values as "true" and "false."

3.7 Implementation Requirements

- ❖ Identifying Problems
- ❖ Picking a tool & building a strategy
- ❖ Assembling Data Sets
- ❖ Building our Model
- ❖ Optimizing, Testing & Deploying Models

CHAPTER 4

Experimental Result and Discussion

4.1 Experimental setup

The setup we employ to evaluate our classification is as follows: In Chapter 3, we compare six approaches. We use LR and five different classifications for the classification. For this, we use Python with Pandas, NumPy, Matplotlib, Seaborn, Sklearn, and NLTK. Then the model was run on our dataset to accomplish our goal.

4.2 Experimental Results

When compared to the other datasets, the predictions made by the training dataset are by far the most reliable. On the other side, we can see that the test data set has a low percentage and is not good at predicting. As a result of testing with six different algorithms, we've determined that the training dataset has high prediction accuracy. We tried six type of model and got the result for both training and test dataset accuracy,precision,recall and fl score

In the table no(3): we can see that six training and test accuracy ,precision, Recall and FL score with all the true value

Table 3: Model Performance (Supervised Machine Learning)

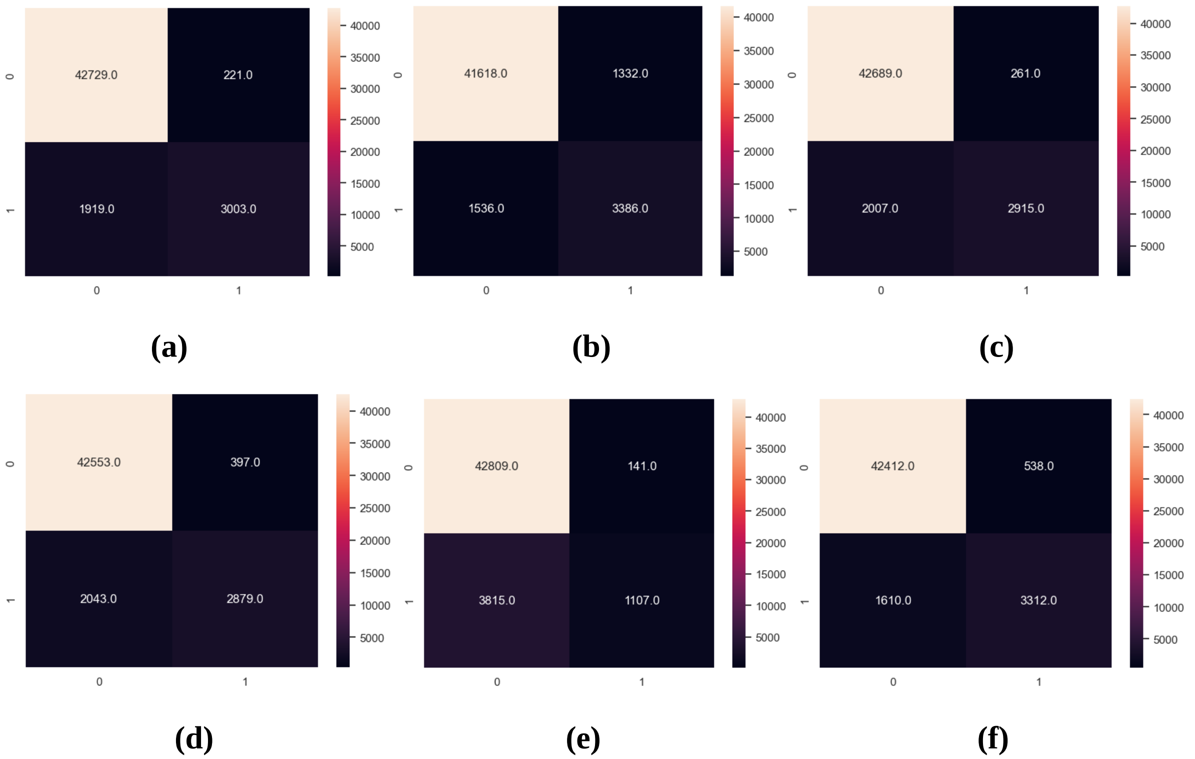
Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	Training 0.95 Test 0.95	0.93	0.61	0.74
Decision Tree Classification	Training 0.99 Test 0.94	0.72	0.69	0.70
Random Forest Classification	Training 0.99 Test 0.95	0.86	0.67	0.76
XGBoost classification	Training 0.96 Test 0.95	0.92	0.59	0.72
AdaBoost Classifier	Training 0.95 Test 0.94	0.88	0.58	0.70
KNN Classification	Training 0.92 Test 0.91	0.89	0.22	0.36

4.3 Experimental Analysis

We obtained 0.95 and 0.95 from the training and test datasets, respectively; the total data from the training dataset is 42729 and 221 bytes; and the total data from the test dataset is 1918 and 3004 bytes. We obtained precision of about 0.96 and 0.93 from the training and test datasets, respectively. According to the report, the recall percentage on the test dataset is low, at 0.61, which is a low value for a prediction, which is why this data set has low accuracy, at 1, whereas the training dataset has all percentages above 0.90, which is why it is at 0. The decision tree was successful in achieving a training accuracy of 0.99 and a test accuracy of 0.94, both of which are considered respectable. The overpowered True class was able to manage a 0.72 percent precision value and a 0.70 value on the f1 score, according to further investigation; however, cracks appear to show where these achievements were accomplished. The fact that this value isn't very close to 1 reveals that the model isn't very good at predicting what will happen in the future. The model is not functioning as well as it should, and it cannot generalize the labels of the truly positive classes to an adequate degree. Having said that, the accuracy of the datasets is quite high. The training dataset accuracy is 0.99, and the test dataset accuracy is 0.95, according to the random forest classification report. In addition, the training tree has 0.96 percent precision and accuracy, while the test tree has 0.86 percent. The dataset also provided us with a fl-score of 0.98 percent for the training dataset and 0.76 percent for the test dataset. As we speak, we've gotten a significant amount of accuracy by using random forward classification. In here, we also see that 1 reveals that the model isn't very good at predicting. The XGBoost classification report shows us that the accuracy of the training dataset is 0.96 and the test accuracy is 0.95. Also, we can see that the test tree has a precise accuracy of 0.92 percent, whereas the training tree has 0.96 percent. Additionally, the dataset was able to provide us with a fl-score of 0.97 percent for the training dataset and 0.72 percent for the test dataset. While we were speaking, we discovered that utilizing random forward classification helped us achieve a large degree of accuracy. In addition, 1 indicates that the model is not particularly accurate when it comes to making predictions. According to the AdaBoost classifier report that we applied to the dataset, we obtained 0.95 percent and 0.94 percent from the training and test datasets, respectively. Although we obtained test data in a similar manner, it is not suitable for prediction, which is why it is 1 and the precision is 0.95 and 0.85 for the training and test datasets, respectively. These two data sets yielded FL scores of 0.97 and 0.70 percent, respectively. We obtained the desired level of accuracy from the dataset by using the AdaBoost classifier. We know from our neighbors that it uses an axis to determine the value and then gives us the accuracy, so we only have 1 axis in this case, and the accuracy for training is 0.92 and for testing is 0.91. In this case, K is the nearest neighbor, which provided precision of 0.92 percent for the training dataset and 0.89 percent for the test dataset. We can see that here the test accuracy is not that good for predicting, which is why it's 1 for the test dataset. We can also see that the fl score is 0.96 percent, which is a good amount of accuracy, but the fl score for the test dataset is 0.36 percent, which is not very good for predicting the comment.

We have classified six types of models that use the confusion matrix to determine the dataset value: true positive, false positive, false negative, and true positive. The goal of this model is to determine the true positive value. As we can see from the result, all the models have successfully classified the true positive value with the maximum dataset value. So the model is accurately showing the correct value, and all the maximum values are true positives.

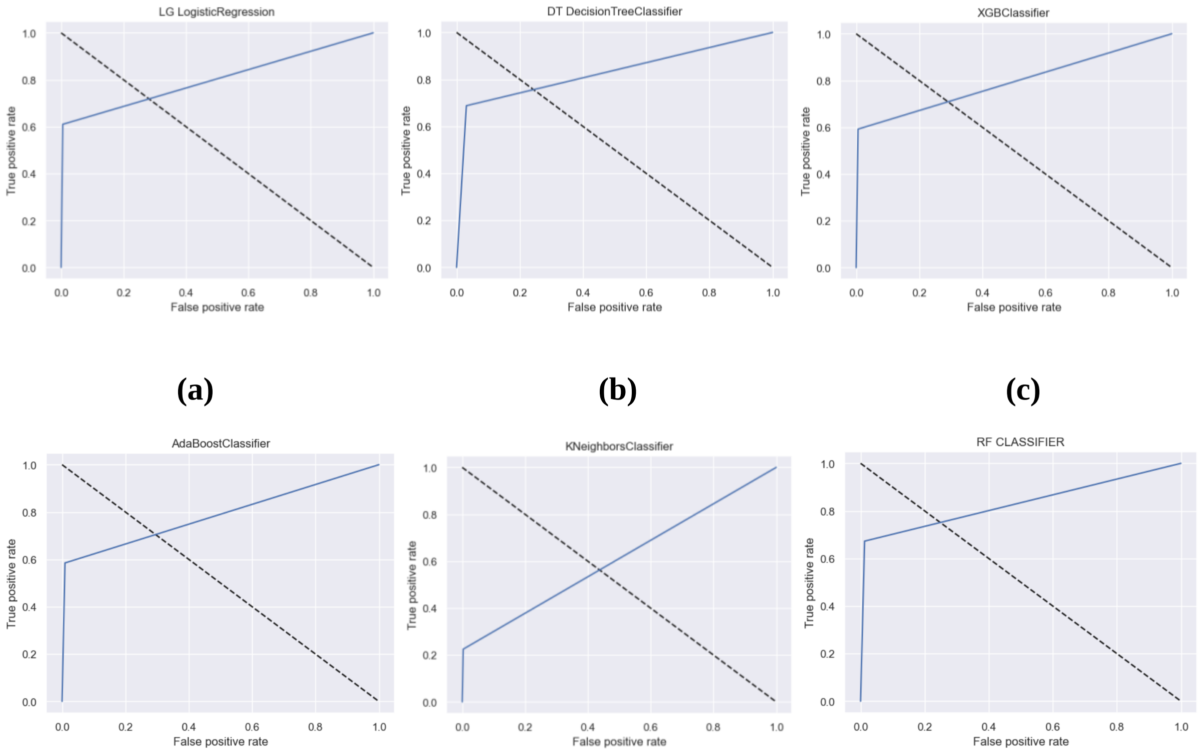
We have also used the plotting graph to see the carving values of all six models. As we can see, the model as a whole is showing a true positive value, but the KNN classification is not. KNN classification uses a different kind of algorithm, which is why it couldn't come up with the right answer while the other five models did. So we can say that the area under the curve is revealed by plotting the graph; the larger the area under the curve, the better the prediction. The curve indicates that the model is performing well for the datasets.



(a) (b) (c)

(d) (e) (f)

Figure 6 : Confusion matrix of (a) Logistic Regression (b) Decision Tree Classification (c) XGBoost classification (d) AdaBoost Classifier (e) KNN Classification (f) Random Forest Classification



(a) **(b)** **(c)**

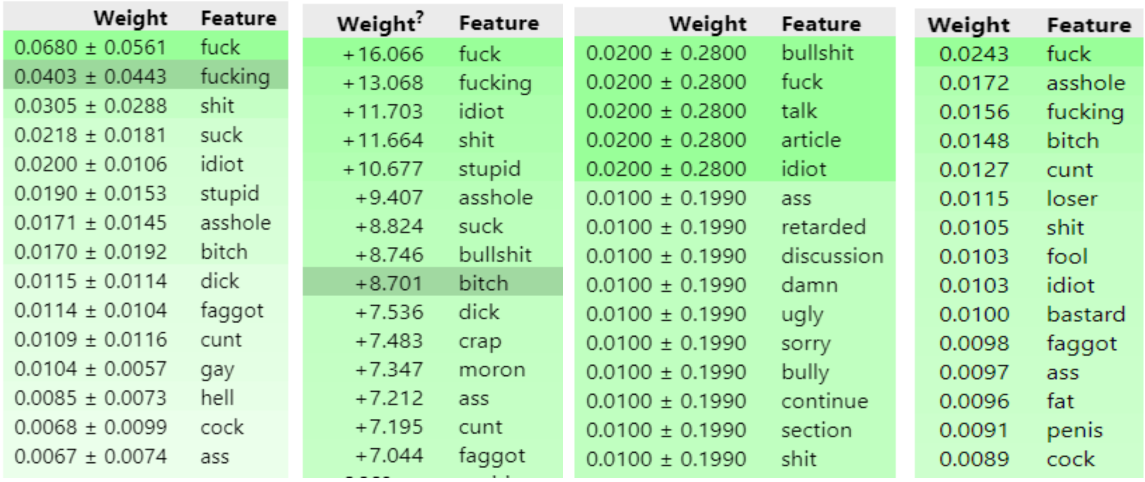
(d) **(e)** **(f)**

Figure 7 : Plotting Graph of (a) Logistic Regression (b) Decision Tree Classification (c) XGBoost classification (d) AdaBoost Classifier (e) KNN Classification (f) Random Forest Classification

4.4 Discussion

From the report we got from Random Forest, Logistic Regression, AdaBoost Classifier, and XGBoost Classifier, we can see the feature that has some toxic words and the weight of that word, which tells us how malignant it is. So, the value of the word is right because the report says that the word's weight score is very toxic.

So, we know that the dataset and model are showing us their real value and have done a perfect job of classifying them.



Weight	Feature
0.0680 ± 0.0561	fuck
0.0403 ± 0.0443	fucking
0.0305 ± 0.0288	shit
0.0218 ± 0.0181	suck
0.0200 ± 0.0106	idiot
0.0190 ± 0.0153	stupid
0.0171 ± 0.0145	asshole
0.0170 ± 0.0192	bitch
0.0115 ± 0.0114	dick
0.0114 ± 0.0104	faggot
0.0109 ± 0.0116	cunt
0.0104 ± 0.0057	gay
0.0085 ± 0.0073	hell
0.0068 ± 0.0099	cock
0.0067 ± 0.0074	ass

Weight?	Feature
+16.066	fuck
+13.068	fucking
+11.703	idiot
+11.664	shit
+10.677	stupid
+9.407	asshole
+8.824	suck
+8.746	bullshit
+8.701	bitch
+7.536	dick
+7.483	crap
+7.347	moron
+7.212	ass
+7.195	cunt
+7.044	faggot

Weight	Feature
0.0200 ± 0.2800	bullshit
0.0200 ± 0.2800	fuck
0.0200 ± 0.2800	talk
0.0200 ± 0.2800	article
0.0200 ± 0.2800	idiot
0.0100 ± 0.1990	ass
0.0100 ± 0.1990	retarded
0.0100 ± 0.1990	discussion
0.0100 ± 0.1990	damn
0.0100 ± 0.1990	ugly
0.0100 ± 0.1990	sorry
0.0100 ± 0.1990	bully
0.0100 ± 0.1990	continue
0.0100 ± 0.1990	section
0.0100 ± 0.1990	shit

Weight	Feature
0.0243	fuck
0.0172	asshole
0.0156	fucking
0.0148	bitch
0.0127	cunt
0.0115	loser
0.0105	shit
0.0103	fool
0.0103	idiot
0.0100	bastard
0.0098	faggot
0.0097	ass
0.0096	fat
0.0091	penis
0.0089	cock

Figure 8 : Features or words that make a comment toxic

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

We have endeavored to create something that will have a beneficial effect on society. By looking at these factors, we may conclude that our society is being negatively impacted by social media and other online platforms that bring people together in some way. This project might help us to bring that capability to fix these issues, and it might bring justice for those who are not using social media to spread hate speech or rude behavior. In the time of matter that people will be rude and thread by their comment and their experience will be bad, and since not everyone we can control by one, so this project might help us to bring that capability to fix these issues.

5.2 Impact on Environment

Therefore, we are aware that the environment we collectively desire on the internet is to have one that is free from danger and filth. However, malicious comments have a significant negative impact on the ecology of the internet. We built the platform using many technologies, including AI and ML. The way that it ought to be used is not how people are using it. In order to make it more careful by now, we can also utilize these techniques to produce a better atmosphere on social media, which will assist us to bring about a safe and neet environment all over the world.

5.3 Ethical Aspects

When we set out to write this work, one of our key goals was to simplify the process of identifying malicious comments inside data sets for the aim of ensuring that social media platforms are not abused by machine. If we are able to accomplish what we set out to do along this road, one of our objectives is to have other people implement this technology so that it can be used for the benefit of the entire planet.

5.4 Sustainability Plan

This research investigated the feasibility of employing malignant comment classification as a teaching environment with the intention of fostering mentalities that are capable of dealing with complexity in the field of sustainability. For the sake of constructing a future that is both livable and sustainable, it is essential to get an understanding of the complex interplay of the world's knowledge. In order for us to be an effective component of this system, we need to train ourselves to continuously evaluate the context in which we find ourselves and alter our worldviews in accordance with the findings of those evaluations. The state of the world compels us to extend our viewpoint and take into account not only the immediate repercussions of our acts but also the more far-reaching ones. An education that is sustainable cannot just concentrate on the acquisition of factual knowledge; rather, it must also develop the sustainability of skills in systems thinking and problem-solving.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication For Future Research

6.1 Summary of the study

As part of this project, we attempted to recognize and categorize abusive and threatening comments made by one person to another by malignant comment classification . We have made an effort to collect data that can be used to identify them with a high degree of accuracy and success, and we have been successful in doing so. Using the train and test model that we constructed, it is possible to detect the problematic phrase with a level of accuracy that is acceptable. Every endeavor was made with the intention of achieving a higher level of precision. The major objective of our investigation was to locate potentially threatening or insulting comments.

6.2 Conclusion

In an effort to improve productivity and produce a summary that is both more accurate and more useful, we had a discussion about a number of the important details that are relevant to the malignant comment classification . In this study, we not only discussed the most significant approaches but also the most important processes that are involved in deep learning. Streamlining the process of identifying harmful comments with the use of supervised learning was the primary focus of our efforts.

6.3 Implication for further study

Over the course of the past few years, advancements in science and technology have made our lives simpler and more efficient. This work has been carried out with a fair amount of precision by us, however it could be improved upon and made more reliable. The utility of the survey model may be improved in the not-too-distant future through the implementation of alternative procedures, the introduction of fresh parameters, and the development of additional features.

Any platform that has suffered damage as a result of comments that are offensive or dangerous can benefit from more research in order to improve their accuracy. We have already constructed a model dataset that contains some offensive words, and the results of our work have been reliable. We are able to leverage these platforms to obtain further information, and then we can use that information to assist the platforms that are being negatively impacted by the cause. We had some difficulty putting together the program and other code that had to do with locating malicious remarks; nonetheless, we attempted so many times, and in the end, we were successful. Future plans call for the incorporation of AI and ML into the affected platforms, as well as the enhancement of accuracy through the use of fresh data sets that may be obtained from those platforms.

References

[1] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 1391–1399.

[2] Coversation AI Team, https://conversationai.github.io

[3] Koppisetti4 1Associate Professor, Dept. of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, A.P.,, India:Identification and Classification of Toxic Comment Using Machine Learning MethodsReceived: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

[4] CS224N: Detecting and Classifying Toxic Comments Kevin Khieu Neha Narwalkkhieu@stanford.edu nnarwal@stanford.edu

[5] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining Word Embeddings for Sentiment Analysis," Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017.

[6] Van Aken, Betty Risch, Julian Krestel, Ralf Löser, Alexander. (2018). Challenges for Toxic Comment Classification: An In-Depth Error Analysis. 10.18653/v1/W18-5105.

[7] Detecting Abusive Comments in Discussion Threads Using Naïve Bayes,Md. Abdul Awal ,Department of Computer Science and Engineering, Khulna University of Engineering & Technology,Khulna 9203, Bangladesh awal.kuet@yahoo.comMd. Shamimur Rahman Department of Computer Science and Engineering, Khulna University of Engineering & Technology, Khulna9203, Bangladesh shamimur052@gmail.com Jakaria Rabbi Department of Computer Science and Engineering, Khulna University of Engineering &Technology, Khulna9203, Bangladesh jakaria.rabbi@yahoo.com 2018 2nd Int. Conf. on Innovations in Science, Engineering and Technology (ICISET) 27-28 October 2018, Chittagong, Bangladesh

[8] Identification and Classification of Toxic Comments on Social Media using Machine Learning Techniques P. A. Ozoh1, A. A. Adigun2 , M. O. Olayiwola3 1,2Department of ICT, Osun State University, Osogbo, Nigeria 3Department of Mathematical Sciences, Osun State University, Osogbo, Nigeria.International Journal of Research and Innovation in Applied Science (IJRIAS) | Volume IV, Issue XI, November 2019|ISSN 2454-6194

[9] Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794

[10] Toxic Comment Detection and Classification,Prinslou Tare ,prinslou@stanford.edu Department of Computer Science Stanford University

[11] Navoneel Chakrabarty1: A Machine Learning Approach to Comment Toxicity Classification ; 1 Jalpaiguri Government Engineering College,Jalpaiguri, West Bengal, India nc2012@cse.jgec.ac.i

[12] Malignant Comment Classification available at <<https://www.kaggle.com/datasets/surekharamireddy/malignant-comment-classification />> last accessed on 06-12-2022 at 12:00 PM.

[13] Multi-task learning for toxic comment classifcation and rationale extraction ,Kiran Babu Nelatoori1 · Hima Bindu Kommanti1,Received: 21 April 2022 / Revised: 29 June 2022 / Accepted: 30 June 2022 © The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022.

[14] KNN Classification Tutorial using Scikit-learn available at << https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn >> last accessed on 08-12-2022 at 12:00 PM.

[15] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146, 2017.

[16] Pete Burnap and Matthew L Williams. Us and them: identifying cyber hate on twitter across multiple protected characteristics. EPJ Data science, 5(1):11, 2016.

Malignant comment classification Using Machine Learning Model

ORIGINALITY REPORT

25%	22%	13%	15%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	cs229.stanford.edu Internet Source	3%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
3	Submitted to Daffodil International University Student Paper	2%
4	Kiran Babu Nelatoori, Hima Bindu Kommanti. "Multi-task learning for toxic comment classification and rationale extraction", Journal of Intelligent Information Systems, 2022 Publication	2%
5	www.rsisinternational.org Internet Source	1%
6	Submitted to Nottingham Trent University Student Paper	1%
7	github.com Internet Source	1%