**CERTIFICATE I**

This is to certify that, the Field Training undergone in Global Engineering Deans Council, India Chapter and Rajarambapu Institute of Technology, Rajaramnagar is a bonafide work carried out by Shradha S. Kumbhar under my guidance. The report is submitted toward the partial fulfillment of two year, full time Post Graduate Program in Computer Science & Engineering.

.....

Guide

.....

Head of Program.

.....

Head of Department

…….....

Director

## CERTIFICATE II

Global Engineering Deans Council, India Chapter and
Rajarambapu Institute of Technology, Rajaramnagar happy to award

Certificate of Completion

to

Shradha Sambhaji Kumbhar of Rajarambapu Institute of Technology,Rajaramnagar for completing the 5 weeks Virtual Internship Program (150 Hours) on Machine Learning using Python Programming organized by Computer Science & Engineering Department from 22nd June 2020 to 22nd July 2020 and secured grade AB .

Dr. N. V. Dharwadkar — HOD
Dr. A. P. Shah — Co-ordinator VIP
Dr. S. K. Patil — Dean Academics
Dr. Mrs. S. S. Kulkarni — Director, RIT, Rajaramnagar
Dr. K. Manivannan — Chairperson, GEDC India Chapter

Certificate ID: RIT-VIP-CSE-04

## ACKNOWLEDGEMENT

This field training report is a result of intense effort of many people to whom I want to thank you for making this reality. Thus I express my deep regards to all those who have offered their assistance, suggestions and training. It is a great pleasure in my presenting training work done at Global Engineering Deans Council, India Chapter and Rajarambapu Institute of Technology, Rajaramnagar. I want to express my deep sense of gratitude, our guide Dr. S. A. Thorat for his inspiring comment and training. I also want to express my deep sense of gratitude, arrangers of virtual internship who give me a chance to complete my training in machine learning. I acknowledges with thanks to faculty members of the CSE Department. Finally, I would like to thanks to all who directly and indirectly help me for same.

# INDEX

# Chapter 1

# Introduction

Nowadays, the age of Internet has changed the way people express their views, opinions. It is now mainly done through blog posts, online forums, product review websites, social media, etc. Nowadays, millions of people are using social network sites like Facebook, Twitter, Google Plus, etc. to express their emotions, opinion and share views about their daily lives. Through the online communities, we get an interactive media where consumers inform and influence others through forums. Social media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc. Moreover, social media provides an opportunity for businesses by giving a platform to connect with their customers for advertising. People mostly depend upon user generated content over online to a great extent for decision making. For e.g. if someone wants to buy a product or wants to use any service, then they firstly look up its reviews online, discuss about it on social media before taking a decision.

The amount of content generated by users is too vast for a normal user to analyze. So there is a need to automate this, various sentiment analysis techniques are widely used. Sentiment analysis (SA) tells user whether the information about the product is satisfactory or not before they buy it. Marketers and firms use this analysis data to understand about their products or services in such a way that it can be offered as per the user's requirements. Textual Information retrieval techniques mainly focus on processing, searching or analyzing the factual data present. Facts have an objective component but, there are some other textual contents which express subjective characteristics. These contents are mainly opinions, sentiments, appraisals, attitudes, and emotions, which form the core of Sentiment Analysis (SA). It offers many challenging opportunities to develop new applications, mainly due to the huge gr0owth of available information on online sources like blogs and social networks. For example, recommendations of items proposed by a recommendation system can be predicted by taking into account considerations such as positive or negative Introduction opinions about those items by making use of SA.

## 1.1Objectives

How can your business benefit from sentiment analysis?

🛡**Protect Your Reputation**
Find negative comments on Twitter and talk directly to their authors before they turn into a PR crisis.

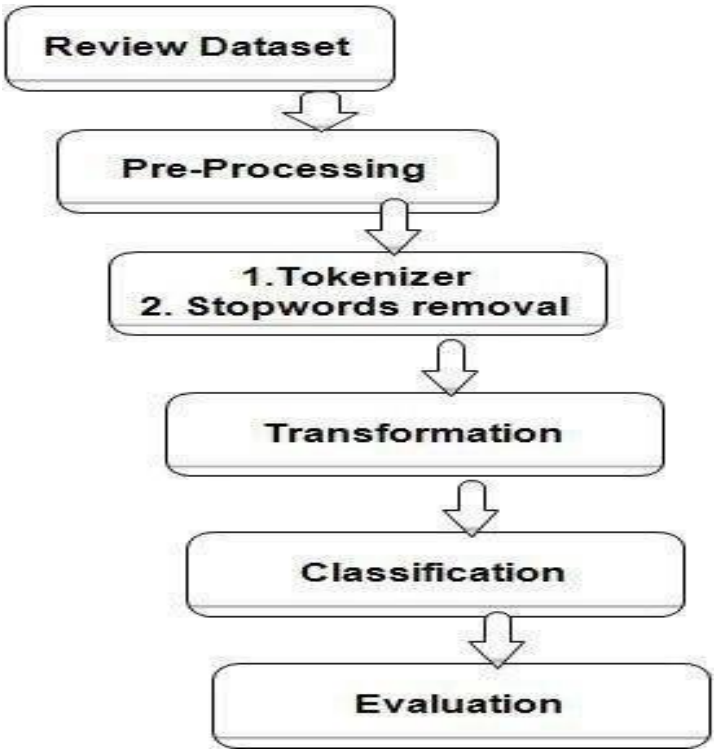
🔎 **Find marketing insights**
check out opinions about your products customers express on Twitter and adjust your offer to meet their needs.

🗣 **Improve your customer service**
Customers talk about products on the web and social media more often than directly to you.

- ❑ Sentiment analysis to determine the attitude of the mass is positive, negative or neutral towards the subject of interest.
- ❑ Apply several algorithms on the model to check out the accuracy.
- ❑ Graphical representation of the results.

**Twitter** is one of the platforms widely used by people to express their opinions and showcase sentiments on various occasions. **Sentiment analysis** is an approach to analyse data and retrieve sentiment that it embodies. The tweet format is very small, which generates a whole new dimension of problems like the use of slang, abbreviations, etc.
This reports on the exploration and preprocessing of data, transforming data into a proper input format and classify the user's perspective via tweets into ***positive (non-racist) and negative (racist)*** by building supervised learning models using Python and NLTK library.

**Quick Workflow**




## 1. Review Dataset

- In this, we have used the two data set one is to test the data and another is to train the dataset
- Comparison of dataset using histogram
- View what's the length of the Tweets in our Train and Test data

## 2. Pre-processing

- Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.
- Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors.
- Data preprocessing is a proven method of resolving such issues

## 3. Transformation

**Tokenization**

It is just the term used to describe the process of converting the normal text strings into a list of tokens i.e. words that we actually want. Sentence tokenize can be used to find the list of sentences, and Word tokenize can be used to find the list of words in strings**.**

**Stemming**

Stemming is the process of reducing inflected (or sometimes derived) words to their stem, base or root form — generally a written word form. For example, if we were to stem the following words: “Stems”, “Stemming”, “Stemmed”, “and Stemtization”, the result would be a single word “stem”.

**Transformation**

Adding the tidy tweets back to our main (merge) data frame. Now we want to see how well the given sentiments are distributed across the training dataset. One way to accomplish this task is by understanding the common words by plotting word clouds.

A word cloud is an image made of words that together resemble a cloudy shape. The clouds give greater prominence to words that appear more frequently in the source text. You can tweak your clouds with different fonts, layouts, and color schemes.

## 4. Classification

Now that we are ready with our pre-modeling stages of the data it’s time to classify them on the sections in which I have prepared the data that is Bag-of-words, TF-IDF, word2vec vectors, and doc2vec vectors. For the implementation we have implemented following algorithms:

1. Logistic Regression
2. Naive Bayes
3. Support Vector Machine

4. Random Forest

5. XGBoost

## 5. Evaluation

For evaluation we apply the several algorithms mentioned above and did the comparative study and analyze which method is efficient for the Sentiment Analysis.

# Chapter 2

# Machine Learning Methods Used

## 2.1 Linear Regression

Logistic Regression is a classification algorithm could help us predict whether the given Tweet is Racist or not. Logistic regression predictions are distinct. We can also view probability scores underlying the model's classifications.

Logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes. There are 3 types of Logistic Regression. Binary, Multi, and Ordinal. Here is the metric evaluation of the Logistic Regression.

## 2.2Naive Bayes

Naive Bayes is a classification technique based on the Bayes' Theorem with an assumption of independence among predictors. Bayes' Theorem provides a way in which an equation is describing the relationship of conditional probabilities of statistical quantities. In Bayesian classification, we're interested in finding the probability of a label given some observed features.

There are three types of Naive Bayes model. Gaussian, Multinomial, and Bernoulli. Gaussian is basically used for classification problems whereas Multinomial is used to get the distinct counts. Bernoulli is used if the feature vectors are binary.

Bernoulli is effective in text classification with a bag of words models where the 1 & 0 are "words occur in the document" and "words do not occur in the document".

## 2.3Support Vector Machine

SVM is a supervised machine learning algorithm which is used for classification and regression problem. In SVM we perform the classification by finding the hyper plane that differentiates between two classes very well. The Kernel trick technique is used to transform data and then based on the data it finds an optimal boundary between the possible outputs.SVM can be optimized using hyper parameters such as Kernels, Regularization,

Gamma, and Margin.SVM works really well with a clear margin of separation and high dimensional spaces.

## 2.4 Random Forest

Random forest is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use the algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests create decision trees on randomly selected data samples, get a prediction from each tree, and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

A random seed is chosen which pulls out at random collection of samples from the training dataset while maintaining the class distribution. With this selected dataset, a random set of attributes from the original dataset is chosen based on user-defined values. All input variables are not considered because of enormous computation and high chances of over fitting.

## 2.5 XGBoost

There are 4 Boosting Machine Learning Algorithms

1. Gradient Boosting Machine
2. Extreme Gradient Boosting Machine
3. Light GBM
4. Cat Boost

Boosting is one such technique that uses the concept of ensemble learning. A boosting algorithm combines multiple simple models to generate the final output. The working procedure of XGBoost is that it combines the predictions from multiple decision trees. All the weak learners in a gradient boosting machine are decision trees. The trees in XGBoost are built sequentially, trying to correct the errors of the previous trees.

XGBoost implements parallel pre-processing at nodes level which makes it faster than GBM. Using the Regularization technique XGBoost prevents over fitting and improves the overall performance.

# Chapter 3

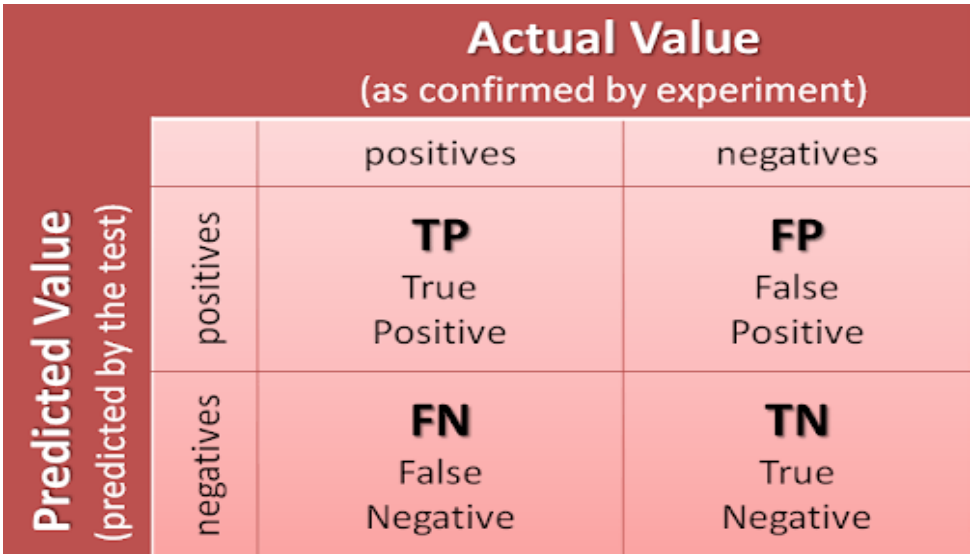
# Experimental Results

## 3.1 Parameters used

But just building the algorithms isn't sufficient. We need to evaluate our algorithms against some criteria. So the criteria that I have considered to measure the performance of the implemented models are Confusion Matrix, Accuracy, AUC (Area under Curve), and F1 Score.

- **Confusion Matrix**

  It is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values and gives us a matrix as output and describes the complete performance of the model. Below is the terminology involved in the Confusion Matrix.




**True Positive:**

Interpretation: You predicted positive and it's true.

You predicted that a Tweet is racist and Tweet actually is.

**True Negative:**

Interpretation: You predicted negative and it's true.

You predicted that a Tweet is not racist and Tweet actually is not.

**False Positive: (Type 1 Error)**

Interpretation: You predicted positive and it's false.

You predicted that a Tweet is racist but Tweet is actually is not.

**False Negative: (Type 2 Error)**

Interpretation: You predicted negative and it's false.

You predicted that a Tweet is not racist but Tweet actually is.

- **Accuracy**

Since we are working on a Classification problem of classifying the tweets (Racist & Non-Racist) Classification Accuracy will be a good evaluator for the models. This is the most commonly used metric to evaluate how well the machine learning model is doing. Accuracy is the ratio of the number of correct predictions made against the total number of predictions made.

Accuracy = (TP + TN)/ (Total Number of Samples)

- **Area under the Curve**

  - Area under Curve (AUC) is one of the most widely used metrics for evaluation. It is used for binary classification problems. AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example.

  - False Positive Rate and True Positive Rate both have values in the range [0, 1] AUC is the area under the curve of plot False Positive Rate vs. True Positive Rate at different points in [0, 1].

  - The area under the ROC curve is often used as a measure of the quality of the classification models. A random classifier has an area under the curve of 0.5,

while AUC for a perfect classifier is equal to 1. In practice, most of the classification models have an AUC between 0.5 and 1

- As evident, AUC has a range of [0, 1]. The greater the value, the better is the performance of our model.

- When a classifier cannot distinguish between the two groups, the area will be equal to 0.5 (the ROC curve will coincide with the diagonal). When there is a perfect separation of the two groups, i.e., no overlapping of the distributions, the area under the ROC curve reaches to 1 (the ROC curve will reach the upper left corner)

- **F1 score**

- F1 Score is the Harmonic Mean between precision and recall. The range for the F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances). High precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model. Mathematically, it can be expressed as :

- F1 Score tries to find the balance between precision and recall.

- Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier.

- Recall: It is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

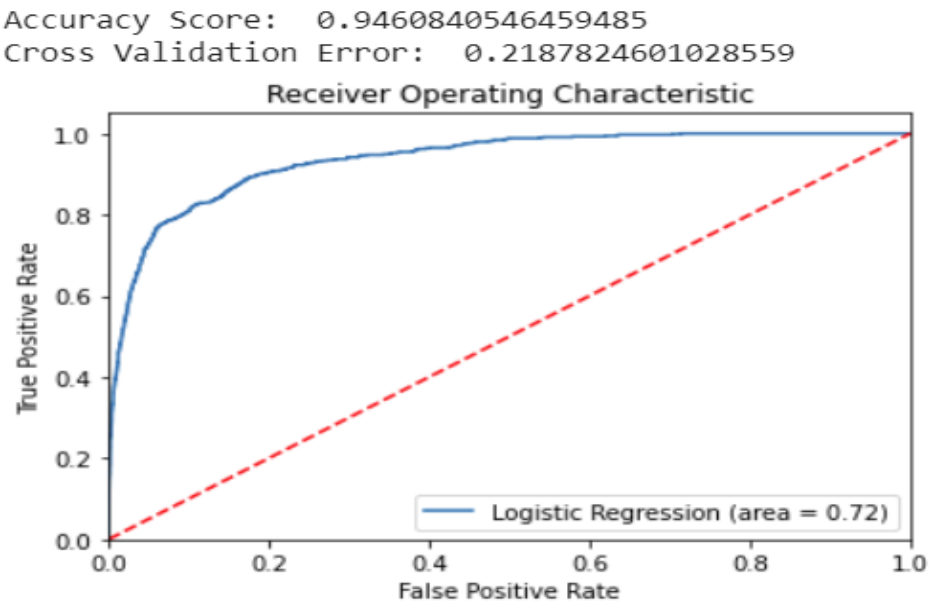
## 3.2 Experimental Results with Logistic Regression

- **Logistic regression** is a statistical **model** that in its basic form uses a **logistic** function to **model** a binary dependent variable, although many more complex extensions exist.
- **Logistic Regression** and Effective Word Score Heuristic.

- **Sentiment Analysis** is a method for judging somebody's **sentiment** or feeling with respect to a specific thing.

In the proposed work, **logistic regression** classification is used as a classifier and unigram as a feature vector.

**Graph:**



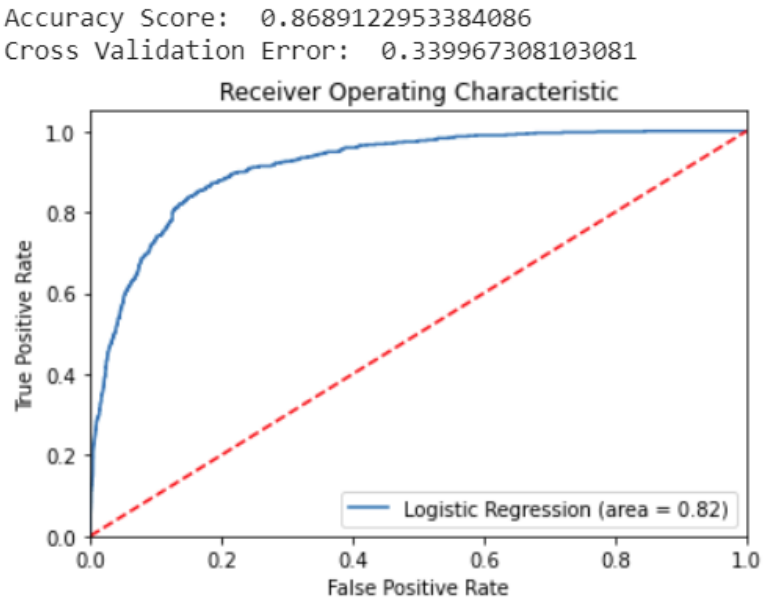

**Classification Report**

| | precision | recall | F1 score | support |
|---|---|---|---|---|
| 0 | 0.97 | 0.97 | 0.97 | 8905 |
| 1 | 0.62 | 0.61 | 0.62 | 684 |
| accuracy | | | 0.95 | 9589 |
| Macro avg | 0.80 | 0.79 | 0.79 | 9589 |
| Weighted avg | 0.95 | 0.95 | 0.95 | 9589 |

## 3.3 Experimental Results with Naïve Bayes

- It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors.
- In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

**Graph:**




**Classification Report**

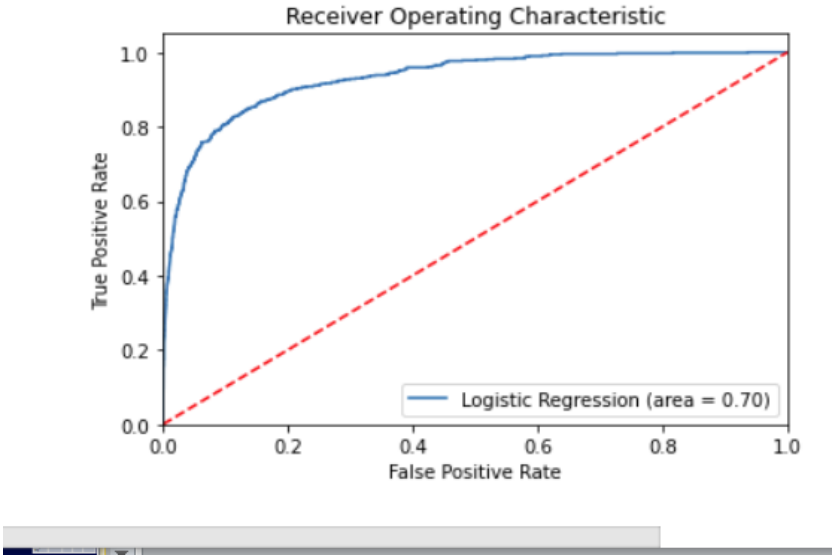
| | precision | recall | F1 score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.88 | 0.93 | 8905 |
| 1 | 0.33 | 0.78 | 0.46 | 684 |
| accuracy | | | 0.87 | 9589 |
| Macro avg | 0.65 | 0.83 | 0.69 | 9589 |
| Weighted avg | 0.93 | 0.87 | 0.89 | 9589 |

## 3.4 Experimental Results with Support Vector Machine

- SVM we perform the classification by finding the hyper-plane that differentiates between two classes very well.
- A support vector machine (**SVM**) is a supervised machine learning model that uses classification algorithms for two-group classification problems.
- After giving an **SVM** model sets of labeled training data for each category, they're able to categorize new text. So you're working on a text classification problem.

**Graph:**

```
[ ] Accuracy Score:  0.9481697778704766
    Cross Validation Error:  0.2207869867361564
```



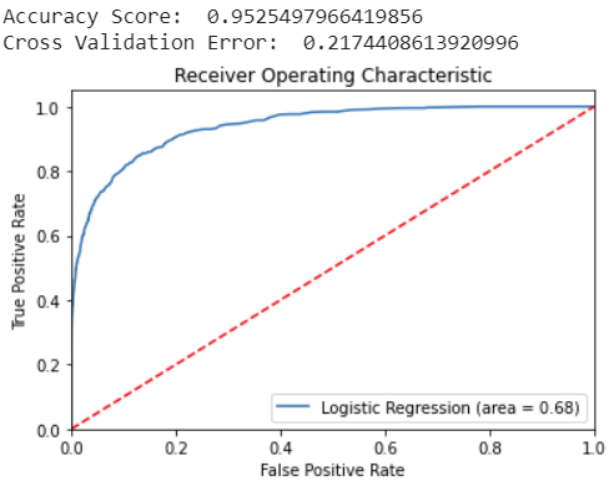

**Classification Report**

| | precision | Recall | F1 score | support |
|---|---|---|---|---|
| 0 | 0.97 | 0.98 | 0.97 | 8905 |
| 1 | 0.65 | 0.58 | 0.62 | 684 |
| accuracy | | | 0.95 | 9589 |
| Macro avg | 0.81 | 0.78 | 0.79 | 9589 |
| Weighted avg | 0.95 | 0.95 | 0.95 | 9589 |

## 3.5 Experimental Results with Random Forest

- Random forest is a supervised learning algorithm. It can be used both for classification and regression.
- **Random forests** or **random decision forests** are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of **decision** trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual.

**Graph:**




**Classification Report**

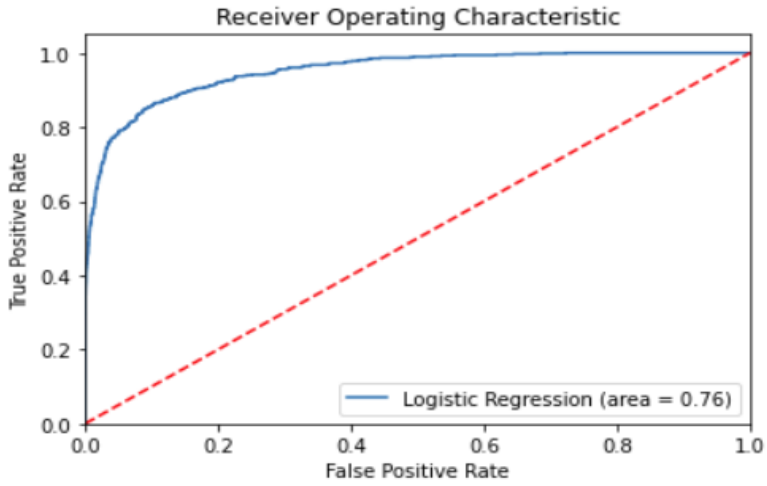
| | precision | Recall | F1 score | support |
|---|---|---|---|---|
| 0 | 0.95 | 1.00 | 0.98 | 8905 |
| 1 | 0.95 | 0.35 | 0.51 | 684 |
| accuracy | | | 0.95 | 9589 |
| Macro avg | 0.95 | 0.68 | 0.74 | 9589 |
| Weighted avg | 0.95 | 0.95 | 0.94 | 9589 |

## 3.6 Experimental Results with XGBoost

- Boosting is one such technique that uses the concept of ensemble learning.
- A boosting algorithm combines multiple simple models to generate the final output.
- The working procedure of XGBoost is that it combines the predictions from multiple decision trees.
- All the weak learners in a gradient boosting machine are decision trees.
- The trees in XGBoost are built sequentially, trying to correct the errors of the previous trees.

**Graph:**

```
Accuracy Score:  0.9594326832829283
Cross Validation Error:  0.1972780145613723
```




**Classification Report**

| | precision | Recall | F1 score | support |
|---|---|---|---|---|
| 0 | 0.96 | 0.99 | 0.98 | 8905 |
| 1 | 0.85 | 0.54 | 0.65 | 684 |
| accuracy | | | 0.96 | 9589 |
| Macro avg | 0.91 | 0.76 | 0.81 | 9589 |
| Weighted avg | 0.96 | 0.96 | 0.95 | 9589 |

## 3.7 Comparative Results

From the above-consolidated analysis, we are able to comprehend that the XGBoost model works pretty well with our data improving the f1 score and accuracy relative to the other machine learning algorithms applied to our model.

On the evaluation of the models like Logistic Regression, SVM, Random Forest, and XGBoost on various features extracted which are Bag of words, Word2Vec, Doc2Vec, and TF_IDF. Considering the evaluation metric of the F1 score, our best performing model is XGBoost with tuned prams applied on the Word2Vec features with an F1 score of 0.66.

**Actual Output**

| | Model_ID | F1Score | Accuracy |
|---|---|---|---|
| 0 | Logistic Regression | 0.621402 | 0.946501 |
| 1 | Naive Bayes | 0.445378 | 0.862342 |
| 2 | Support Vector Classifier | 0.610510 | 0.947440 |
| 3 | Random Forest CLassifier | 0.502697 | 0.951924 |
| 4 | XGBOOST | 0.662555 | 0.960058 |

# Chapter 4

# Conclusion

In this project, we used multiple training models to make a prediction and tried to find the most appropriate model on our dataset. During the process, we learned the analyzing steps and training models in NLTK and sci-kit-learn library which are helpful and easy to use. With different victories, TFIDF, and Count Victories, it will generate different results. For example, we can improve the model by removing words in very low or high frequency by TFIDF. It will make the prediction model filter outliers and improve the accuracy of the model.

## Future Scope

As we have worked for several algorithms for sentiment analysis we can use this also for another several social platforms like Google, Facebook in case of exchanging ideas and opinions. As we all are aware in this grown social media one bad word or statement can raise the controversy or hurt anyone sentiment it results into bad review of that particular platform also. For such circumstances this sentiment analysis can help us to detect the racist content and block it.

# Chapter 5

# References


- https://medium.com/analytics-vidhya/twitter-sentiment-analysis-b9a12dbb2043
- https://www.kaggle.com/arkhoshghalb/twitter-sentiment-analysis-hatred-speech/data
- International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016 ,Sentiment Analysis of Twitter Data: A Survey of Techniques