# Offensive Language Detection Using Machine Learning

**By**

Abid Ullah (CS091182006)

Waqar Ahmad (CS091182049)

Aziz Ur Rehman (CS091182014)

**Supervisor**

Ma’am Noor Asmat

Institute of Computing

KUST, Kohat

**Institute of Computing**

**Kohat University of Science and Technology, Kohat-26000**

**Khyber Pakhtunkhwa, Pakistan**

**(May), (2022)**

# Offensive Language Detection

# Using Machine Learning

Abid Ullah (CS091182006)

A thesis submitted in partial fulfillment of the requirements for the degree of

BS (Computer Science).

Thesis Supervisor: **Noor Asmat**

Thesis Supervisor's Signature:

____________________________________

FYP Coordinator's Signature:

____________________________________

Director's Signature: ____________________________________

**Institute of Computing,**

**Kohat University of Science and Technology, Kohat-26000**

**Khyber Pakhtunkhwa, Pakistan**

**(May), (2022)**

# Declaration

I certify that this project titled "**Offensive Language Detection Using Machine Learning***"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources has been properly acknowledged / referred.

Signature of Student:

______________________________

# Plagiarism Certificate (Turnitin Report)

This thesis has been checked for Plagiarism. Turnitin report endorsed by Supervisor is attached at the end of thesis.

Signature of Student

Signature of Supervisor

# Copyright Statement

# Acknowledgments

First, I would like to thank my supervisor Ma’am Noor Asmat, teachers of computer Science department from Kohat University of science and technology Hangu Campus. Ma’am Noor Asmat offered her endless advice and inspiration during this thesis.

I

Acknowledge her for the efficient supervision and countless struggle, she put into Teaching me in the technical field.

Finally, I must definite my very deep appreciation to my parents and to my all Friends for given that me with reliable support and endless inspiration during my Years of study and through the process of studying and writing this thesis.  This Achievement would not have been imaginable without them.  Thank you so much.

The computer Science department kindly supported the Kohat university of science and technology, KUST Hangu campus. I am grateful to all my friends.

To my parents for their unconditional affection and for always being there for me.

To my family and friends for their support and encouragement.

*“This document is purely dedicated to my beloved daughter*

*“AJWA ABID” ”*

# Abstract

The main object of this thesis is to build automatic system for detecting offensive language on social media (post, or comments etc.) by using machine learning and deep learning methods/techniques. This project consists of detecting offensive language and classifying the type and target of the offense, profanity or insult etc.

For this project, we are going to use offensive language identification dataset (OLID). With the exponential increase of social media users, cyber bullying has been emerged as a form of bullying through electronic messages. Social networks provide a rich environment for bullies to uses these networks as vulnerable to attacks against victims. Given the consequences of cyber bullying on victims, it is necessary to find suitable actions to detect and prevent it. Machine learning can be helpful to detect language patterns of the bullies and hence can generate a model to automatically detect cyber-bullying actions. This project proposes a supervised machine learning approach for detecting and preventing cyber-bullying. Several classifiers are used to train and recognize bullying actions.

Hate Talk is rude or abusing talk against a group of folk, established traits to a degree race, ritual, preference of sexual partner and masculine. It is banned and established the current charter in the United States of America and the EU, nevertheless the WWW and public news fashioned it likely to spread antagonism surely, fast and anonymously. The large dossier created through friendly publishing terraces demands the incident of a direct mechanical model to discover aforementioned content. We study the conduct of various theme likeness methods and categorization algorithms, addressing to capably handle the connected to the internet bad language bias task. We try miscellaneous likeness methods in the way that Bag of Conversation (Big of Words), discussion and individuality Bag of n-grams, emotions, arrangement and alphabet study, discussion embed-dings and n-gram graphs. Also, we test diversified categorization algorithms: Childlike Bayes, Logistic Reversion, Haphazard Jungles, K-Most forthcoming Neighbors and Artificial Affecting animate nerve organs Networks.

Our aim is to search out judge likeness and categorization algorithms concerning their gift to conduct in the Hate Talk discovery task. Additionally, we focal point the serviceability of n-gram graphs (NGGs) as an adept, low dimensional idea likeness that builds correspondence headings that perform amount to deep looks accompanying important offering to the categorization results. Other than the twofold categorization experiments, we furthermore test our form in multi-class categorization experiments on bad language bias tasks. Our results showed that NGGs are educational and rich in appearance - in spite of being depicted by headings accompanying ranges prepared the number of likely classes - operating marginally worse than the Bag of Dispute and discussion embedding, that are in contrast form by extreme-spatial likenesses. We moreover kill mathematical tests, to test whether NGGs have an important gift to the results. The tests not only shows that NGGs are important face concerning the categorization result, but again that the consolidation of the three best operating countenance (Big of Words, NGGs and discussion embedding) achieves high-quality categorization accomplishment, accompanying the use of the staying content likenesses flexible run-down results. Decisively, the categorization treasure choice appears inferior, because mathematical results for all the proven algorithms are comparable.

**Key Words:** ***offensive language detection, NLP, Machine learning, hate full speech***

# Table of Content

# List of Figures

# List of Tables

# CHAPTER 1: INTRODUCTION

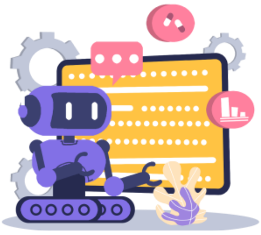


***Figure 1 Automated Content Moderation***

This introduction chapter mainly focuses on four sections. Firstly we will discuss about the motivation behind choosing this particular topic for our AI research project. Here we briefly explain the importance and the major societal impact of choosing this topic and making the contribution in the field. Secondly we will give a brief executive so many about our project describing about the main features along with the functional requirements. Next we will highlight our main contributions achieved in this particular problem area by explaining about the research objectives. Lastly, we present the structure of this thesis report representing what content is there in which chapter and corresponding sections.

Connected to the internet, public news plays a veritably main role in our social relations, which influence our use together, certain negative habits. It specifies society programs to ideas accompanying each one connected to the internet, place folk may be knowledgeable about a great deal of disclosure while just positioned at home. It is veritably near for nations to share many charming, fun, and educational significance actually. Still, there are numerous cases of public addresses named detest addresses that express hate or assure intensity, that cause variable points of detriment to effects or arrangements.

## 1.1 Definition of Hate Speech

The definition of offensive language or hate speech is neither universally accepted nor are individual facets of the definition completely agree upon, Hate speech define by different people and different organization in different way the following show some popular definition among them.

1) Encyclopedia of the American Constitution: Offensive language is the language that attacks an individual or group on the basis of attributes such is ethnic, religion, race sex, origin, or gender.[1]

2) Facebook: Face says that we call everything is offensive if it direct attack on people based on what we call protected properties.---- ethnic, religion, race sex, origin, or gender and disability. Facebook also provide some protections for immigration status.[2]

3) de Gilbert et al: Hate speech is deliberate attack towards a specific group or individuals motivated by aspects of the group’s identity[3]

*Figure 2 Share love not hate*

## 1.2 Motivation

Increase in the usage of social media sites likes Facebook or Twitter have given the crowd a great platform to express their opinions/feelings for the individual, groups of people or events happening around them in society. This digital media has become a great resource to share the information and also gives the full freedom of speech to everyone on the social platforms.

With the gaining popularity of these platforms. Everything has advantages and disadvantages there also comes the negative part along with its benefits. This feature of the social media to express something openly to the world have created the major problems for these online businesses and organization and negatively impacted the well-being of the societal decorum. There are increasing cases of the abuse or offense on the social media like Hate speech, Cyber-bullying, Aggression or general Profanity. It is very much important to understand that this behavior can not only immensely affect the life of an individual or a group but could be suicidal in some cases; adversely hampering the mental health of the victim/s.

This increasing negative situation on the internet has created a huge demand for these social media platforms to undertake the task of detecting or removing the objectionable content and taking the appropriate action which can prevent the situation becoming worse and making these platforms peaceful for everyone. This task of detecting the offensive content can be performed my human moderators manually, but it is both practically infeasible as well as time consuming and very much costly because of the amount of the data generated on these social media platforms its generate huge amount of data in short period of time, therefore there is a need to fill this gap. Numerous studies in the past as evident by the following chapter 2 tries to tackle this problem and gap by leveraging the technologies like Artificial Intelligence, Machine Learning and Natural Language Processing and natural language toolkit. They are able to build the models that can efficiently detect these types of offensive content so that the appropriate action could be taken as quickly as possible to protect user of these platforms.

Last but not least, considering the importance and the sensitivity of this particular problem in the today's digital world there is, still a lot of emerging further scope in tackling and improvise on the previous work done in the field with the help of these new age AI and NLP techniques. Therefore we aim to contribute in this particular direction. The main motive behind our project is that currently major social platform invest $124 M dollar to solve this issue but still there is gap and the problem were not solved yet.

## 1.3 Executive Summary

In this part of executive summary, we will try to summarize the work carried out in all the chapters of our project in some details. We will start off with some Introduction about the project area that will describe the motivation as well as in the project report to take this particular topic. Then we will briefly explain what we have done in the Background chapter to discuss about the extensive review and analysis of the previous work related to our topic (offensive language detection). Then we focus on the architecture and implementation approach that we proposed in this thesis. Finally we conclude this summary by the contributions achieved in this problem area of Offensive Language Detection.

The task that we have taken for this research project is to detect whether a given tweet is offensive or not with the help of NLP techniques along with ML and DL specifically using of NN concepts. The main motivation for choosing this topic is to develop an efficient system to detect offensive content on the social media platform which can be removed afterwards by the human moderators. This will prevent the unease and negativity that may spread in society by using these platforms. We kept people or any community groups in mind.

The background chapter contains all the technical concepts related to AI, ML, DL, NLP, and specifically NLTK which we have used for this project. We have included these details so that it becomes relatively easy for a person who has limited knowledge about the subject to progress towards the later chapters and sections of this document. In this chapter we critically analyzed about fifteen academic research papers that revolve around our project topic. This study helped us to form an appropriate research direction and identifications of the important challenges in the field. This chapter also helped us to come up with the relevant dataset for the problem that is OLID 2019 dataset it is new and many researcher using this dataset because it provide great deal about offensive language.

Further, we focused more on the problem description and also mentioned the important features and the distribution of OLID dataset and preprocess it. As this dataset is the recent one and not much work has been done on it, we laid down some objectives that we can achieve in order to make the significant contribution after the baseline results on this dataset. We divided our proposed solution methodology into mainly five steps. These were Text Pre-processing, Feature Extraction, word embedding, then we develop model not only the one we used four model of ML. at the last we evaluate our model by using Model Evaluation method like confusion matrix etc. In this chapter, we presented our results for the top performing techniques in each of the steps of this NLP pipeline.

As our main contribution to this field, we contributed by making the training dataset balanced using the Random Under-sampling technique that worked best for our problem. Moreover from our comparative analysis of various Machine learning and Deep learning algorithms, SVM model performed the best and achieved the Macro F1 score of 0.99 on this OLID (Offensive Language Identification Dataset 2019) dataset. We also discus about stream_lit module of python language for showing the result in user Interface,  Therefore we made the small improvement in the field when compared with the previous work on this dataset and opened up a new pace to work more on this novel dataset.

## 1.4 Contribution

1. For this studies assignment we researched at the latest OLID short for Offensive Language Identification 2019 dataset which taken into consideration the trouble because the complete with the aid of using taking all of the offense sorts into the account. As the dataset is new and now no longer a good deal paintings has been accomplished on this, consequently to perform our studies to its complete volume and discover new insights the dataset; we accomplished an

intensive comparative evaluation of diverse Feature Extraction Mechanisms like bag of words and tokenization etc., and Model Building Algorithms.

2. We additionally contributed in the direction of mitigating the imbalance with inside the education twitter and Facebook information with the assist of Random Under-sampling Technique.

3. Finally we evolved an exceptional device primarily based totally on nation of the artwork deep gaining knowledge of set of rules SVM, decision tree, KNN and Logistic regration to come across whether a given tweet is offensive or not. We achieve the best Macro F1 rating of 0.99 by SVM which simply outperformed the overall performance of preceding work [1] at the dataset.

### 1.4.1 Visual Aspect

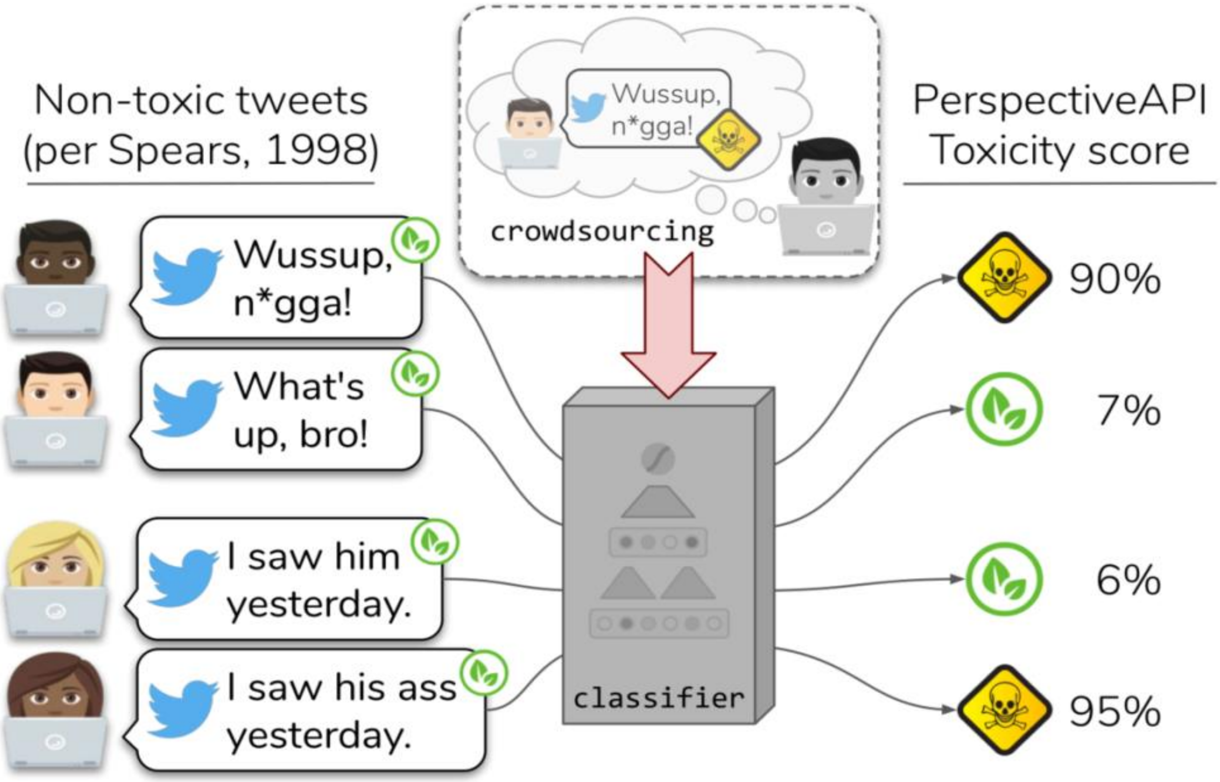


*Figure 3 ML Classifier*

# CHAPTER 2: BACKGROUND

In this chapter we will discuss about the thematic area within Computer Technology In which we are going to explain the basic topic of the project along with the concrete areas and the main area of the Computer Science under which our project falls. Furthermore we deals with project scope which explains all the basics of the topics and the areas explains in the previous section which give us short technical overview of the main domain knowledge to a person who knows nothing or little about the field. Next we move on the literature review of the thematic area which explains the previous work done related to our project topic in terms of motivation and challenges, dataset involved in the previous work, existing methodology, result and evaluation of the models . In the next section, the art focus on critically analyzing the previous works. This will build strong basis to contribute to this specific problem area.

Our research project is about detecting offensive language on social media platforms such as Twitter or Facebook. This problem is very serious and growing. Mostly on these sites, so our goal is to provide solutions to bring them back to normal. Abusive posts using algorithmic techniques in machine learning. This Can effectively reduce the negative impact of these posts on the individual or Group.

We did this work in two versions so we are going to explain both of them one by one in our thesis.

## 2.1 Thematic Area under Computer Technology

The following is thematic areas under computer technology that we are going to learn and implement in our project.

### 2.1.1 Project description

This research study is about the detection of offensive language on the networking sits like Twitter, Facebook, Quora, blogger, and other site in which individuals share their ideas. The problem of hate speech or offensive language is very serious/danger and increasing tremendously of these sites. There for we aim to give perfect solution for normalizing these offensive and hateful posts and comments using a systematic algorithmic methods in ML. This can overcome the negative effect of the posts on the individuals or group of people and also help different organizations which are suffering from such type of content.

### 2.1.2 Project Area

The area under which our project falls is AI: which enable machine to act like human, think like human and mimic all the behaviors of the humans and its sub branch NLP: Which have two parts: NLU stand for natural language understanding and NLG which stand for natural language generating, understand how the human language such as text or speech by the machines in similar way like human beings. We are mainly focusing on NLU which gives machines, the ability to understand the human text (Offensive or not, or hateful or not) and automate the process of identification and removing form all the networking site without human interactions. We will discuss all these terminologies later in this document.

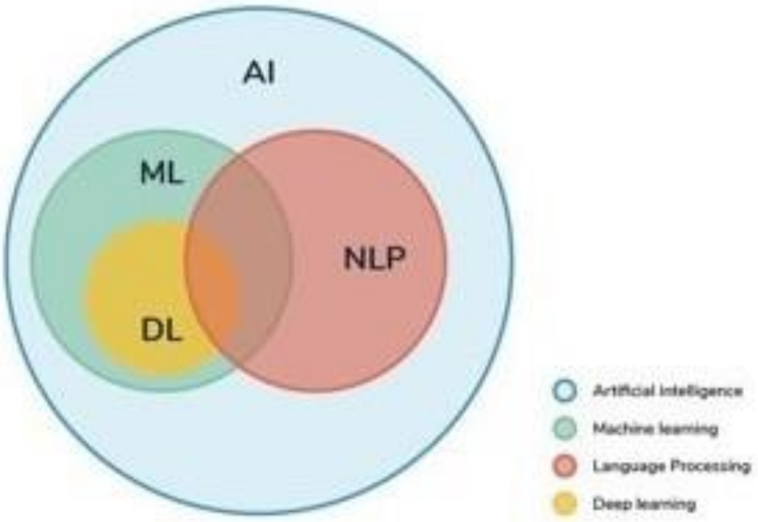


*Figure 4 Different Areas of AI which is the main part of our project*

## 2.2 Project Scope

In this section we are going to discuss the main areas on which we will focus to handle the related things of our project we called this section is project scope it create sense that what we want to achieve and learn them to help humanity and save and protect lives of people using social networking platforms that are discuss earlier. We divide this section into different part like AI, ML, NLP, DL, and some Python libraries and last we will show summery of project scope in the form of slide.

### 2.2.1 Artificial Intelligence

Everyone knows the importance of AI in the recent race of technologies because it helps organization and business a lot. Let's define what AI actually about? AI is define as the ability to enable computer system to automate the process of intellectual tasks that normally require human intelligence , like voice recognition , speech recognition , language translation , understanding and generating language, decision-making and visual perception The AI consists of ML, DL, and many more that don't involve any learning. Which show in figure 7 below?

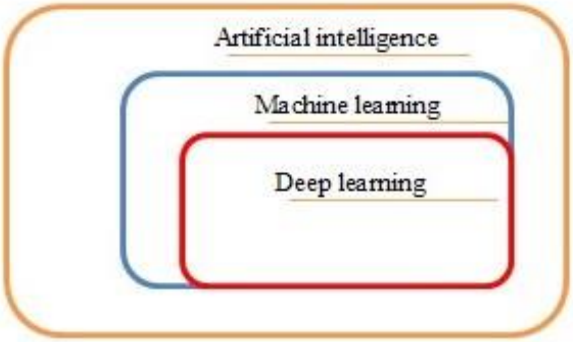


*Figure 5 Sub branches of AI*

### 2.2.2 Basic Process flow of Machine learning

Basic Process flow of Machine learning is shown in the Figure 9 below.

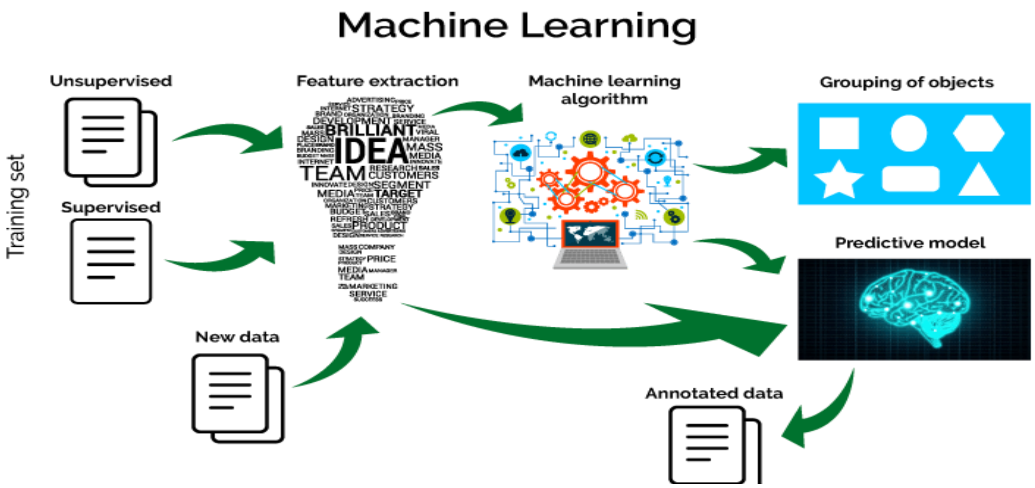


*Figure 6 basic Data flow diagram of ML*

### 2.2.3 Machine Learning

Machine learning is the branch of AI. ML is define as: “Machine learning is the science of enabling computer system to learn from data and act like human do, and improve learning over time in autonomous fashion, by giving them data and information in the form of observation and real world interaction” (Arthur Samuel, 2016)

It is one of the most important application areas of the AI. It needs data to train a model upon it. And by dividing that data in to three set: training data, testing data and validating data.

It has three types:

- Supervised
- Unsupervised
- Reinforcement

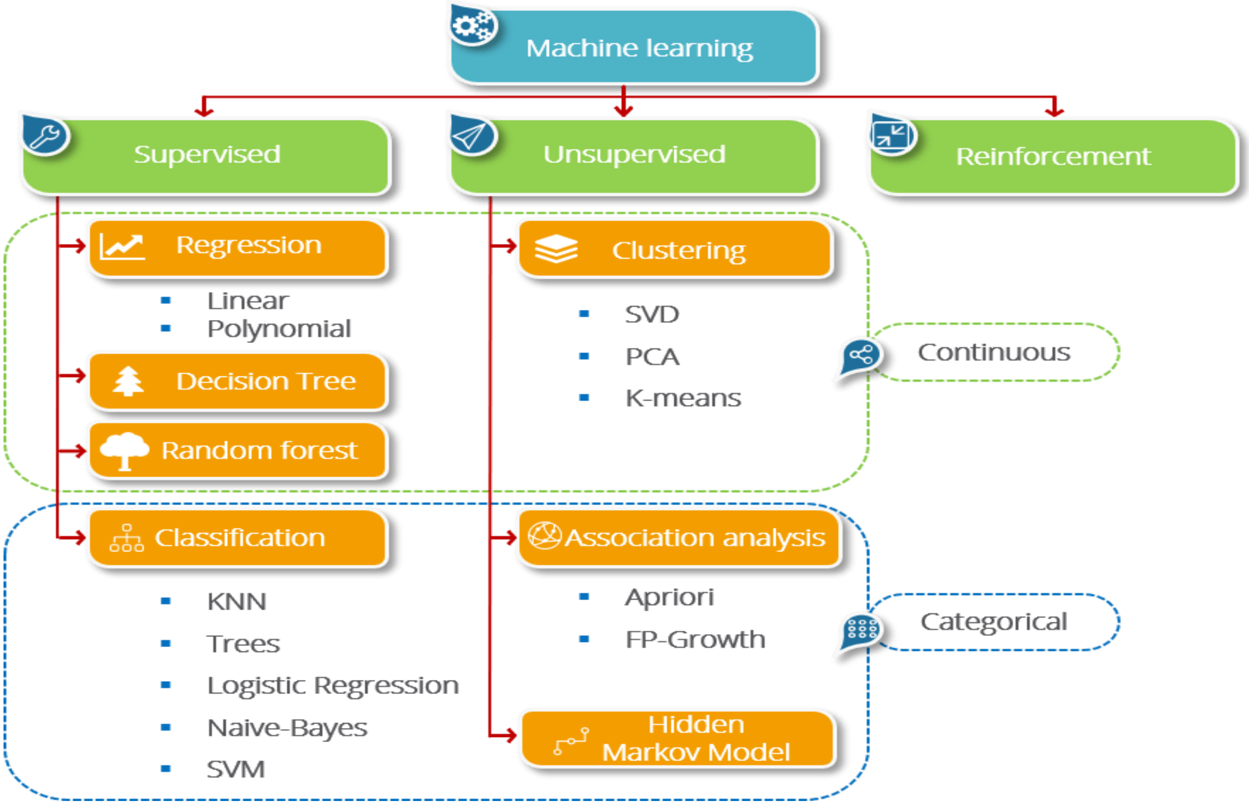


*Figure 7 Types of Machine learning*

In supervise learning method human will provide date with appropriate liable, unsupervised learning method has no prior liable it is use for classification problems and reinforcement method is built upon the reword based learning. It is work on basis of agent and environment. In our case we use supervised learning we will give data with liable such that it offensive and not offensive. Figure 8 shows difference between Traditional modeling and Ml modeling.

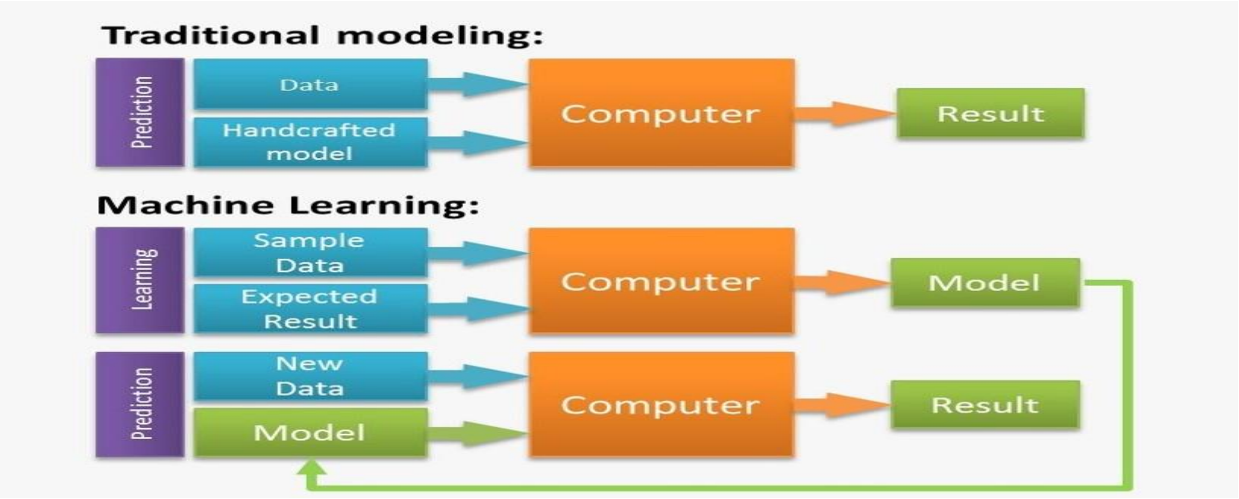


*Figure 8 Difference b/w ML modeling and Traditional modeling*

### 2.2.4 Deep Learning

Deep learning is the sub field of Machine learning. Most advance topic of AI. It can be define as the branch of machine learning that is totally inspired by the biological structure and function of the brain of human containing of billions of neurons that are also called Deep Neural Network because they consist of many layers of neurons which are able to recognize patterns from the raw input data and perform the decision making in the similar way as human brains performs.  DNN learns the hidden insight or pattern from the data through the method called Back-propagation. We have three types of layers in in ANN. ANN is the technique of DNN as DNN has many techniques.

- **Input Layer**

  All the neurons in this layer are connected to each of the feature/property in the dataset. The values of each variable/instance are the inputs to this layer.

- **Hidden Layer**

  In this layer inputs values are multiplied by weights to get the pre-activation outputs then we pass the pre-activated outputs to the activation function which are called sigmoid or logistic function.

- **Output Layer**

This layer will give us the final output like machine learning deep learning has also three types: supervise, unsupervised and reinforcement learning method. The basic structure is shown in the below figure.

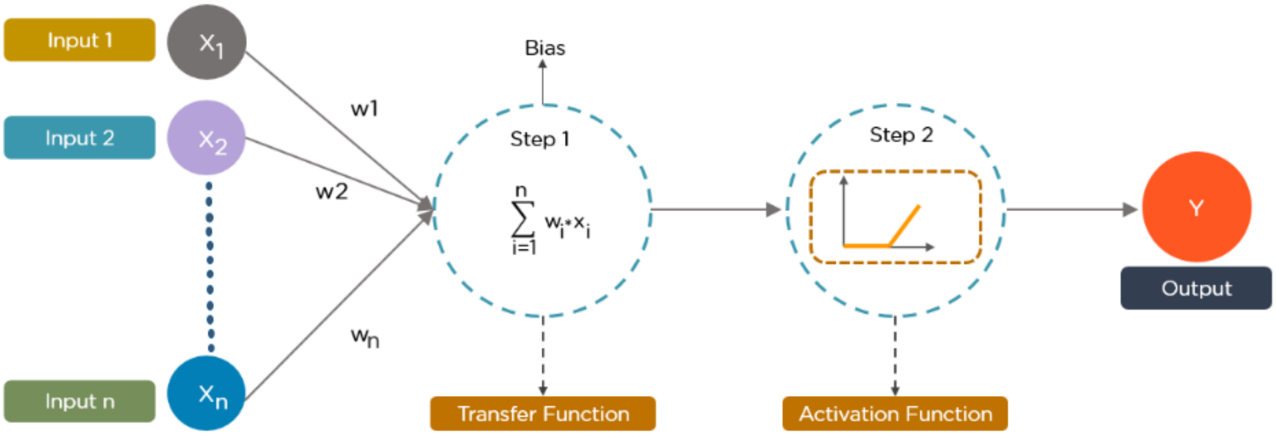


*Figure 9 Simple Artificial Neural Network*

### 2.2.5 Natural Language processing

NLP play an important role in our daily life. Human communicate in the form of language to one another it may be in the form of text or speech. Now if we want to make interactions between human and machine (computers), then machine should know to understand the natural language and to response back to humans it again needs to generate natural language. **NLP is all about to enable computers to learn process and manipulate natural languages to interact with humans.** As our project is about classification of the language in two or more than two groups like offensive, not offensive we need our system to understand NLP [4].

We will explain it in detail in the next section of this document.

## 2.3 A review of Thematic Area

In this part "A review of thematic area" we would analyzing and extracting meaningful information from the few academic papers the we read for the process of offensive language detection in social networking sites using machine learning. We would summaries it in clear way. We will also discuss the common and different points among these papers it may help us to extract the basic structure and the context of the problem solution.

Paper [5] one of the first important tasks was to detect cyber bullying on social media with the help of NLP. This issue has been studied before. Mostly by psychologists and social scientists. The authors thought that anonymity and Lack of meaningful oversight has led the subject to focus more on computer national linguistics. He further added that appropriate action should be taken against the culprits. He further added that appropriate action should be taken against the culprits. The tragic consequences of bullying on social media platforms can be prevented through NLP. The authors modeled this issue in 2 parts. The first step is to determine if it is given. Comment is sensitive or does not make it a binary ranking function. And if there is a comment. Sensitive then categorize it as comments that revolve around a single area of sexuality, the culture or intelligence of the race that made it a multi-class classification sub-task.

Very strong baseline paper [6] aimed at tackling the serious problem of cyber bullying. Using NLP techniques on social media platforms to further investigate. The same domain can be found in the future with the help of the NLP community.

In addition, they developed bullying on social platforms. The 4 major NLP tasks include text classification, role labeling, emotion analysis and Topic modeling. He described the signs of bullying as the posts of people who they are victims of bullying. The purpose of Sub-Task A was to differentiate between bullying. Episodes with non-bullying traces in dataset examples. Sub-task B was dealt with. The roles of the tweets refer to the character of the author and the character of the person along with the labeling. The work of classifying binary text can be thought of. The second thing in this sub-work is this. More sequentially to tag the characters in one of the 5 categories that is to be accused. (A), Bully (B), Reporter (R), Hunter (V) and others (0). Sub-task C was concerned. To classify the emotions of the bullying incidents to understand their motivation. This Task was a binary task of classifying tweets as teasing or not. Sub was related to Task D. With hidden content modeling that brings out the main topics for better understanding. Signs of cyber- bullying.

In [7] the authors felt the need to automate and improve the detection work. Hateful content on social media. The authors make it clear that the previous work has only investigated bullying from individual comments. They strictly speaking, taking user information and profile information can definitely be obtained. Improved model for sensitive material feature.

The writers' motivation for the work [8] stemmed from the murder of the drummer. Le rugby in Woolwich, London, UK Said that he starts it immediately. There is a good chance that hate speech will spread online after such incidents platforms. They formulated this problem as a function of predictive binary classification. Whether the tweet is hateful / hostile or not, race, race or Religion. He took his features from the context of each tweet and experimented. Voted pairs with different probabilities, principle-based, and locally-based classifications the meta-classifier believes that their partnership is closely linked to the effective. Decision making and policy making.

In [9], the authors raised the issue of detecting hate speech using social media User Comments; the growing number of cases has encouraged his work. Hate on online platforms leads to decline in online business and user. Experience their purpose is to create a lesser representation of comments using them. The neural language models they will be sending to the ranking algorithm. They solved the issue of high dimension and sparsely due to previous work of BOW. The model and its purpose are to obtain accurate predictive models for the domain.

In 2016, the authors [10] were the first to focus on the issue of detection. Aggressive language instead of focusing only on specific types on online platforms Misuse such as cyber bullying or hate speech as is evident in recent years. They brought Draw attention to the flaws of regular expressions and blacklists. Matters of hate speech that is more subtle and less obvious than usual. They perform better than a deep learning method to build a prediction. A new corpus of models and user comments also made it unique of its kind.

In [11], the authors challenged the separation of hate speech from other forms failed to distinguish between offensive content as previously used lexical methods. Different types of crime. It is especially important to identify hate speech correctly. From other offensive materials because both have different implications. Societies and individuals. He worked with Twitter Data, which was labeled 3. Categories: hate speech, offensive language, and neither of which he trained a multi-class classifier. Background [11] in [12] the authors exceeded the basic challenge of distinguishing hate speech from [11]. From common insults or crimes to coming up with a literal basis for it.

# CHAPTER 3: Requirements &Proposed Solution

In this chapter we deal with Requirements and proposed system for offensive language detection. First, I will discuss all the problems involve in previous work next, we will discuss all the ML algorithms that we used in our project which include Logistic Regression, K- nearest neighbor, SVM, Random Forest, Decision Tree, and Naïve byes. After that we will highlight our proposed solution.

## 3.1 Problems in Previous Approaches

One of the most and unpredictable problem was the behavior of offensive language people change.

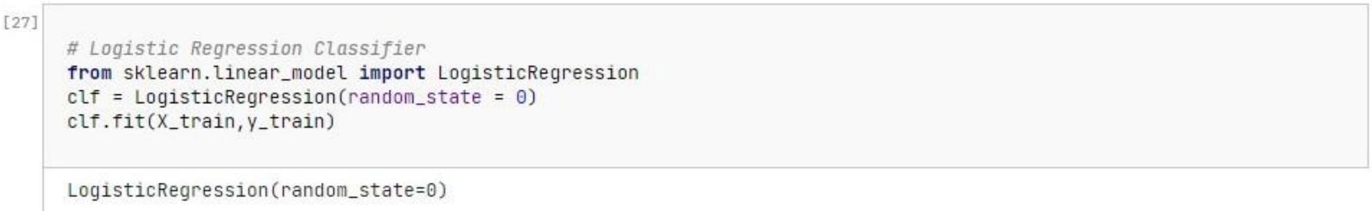
## 3.2 Algorithms Used in Our Project

To solve the said problems we used the following Machine learning algorithms.

### 3.2.1 Logistic Regression

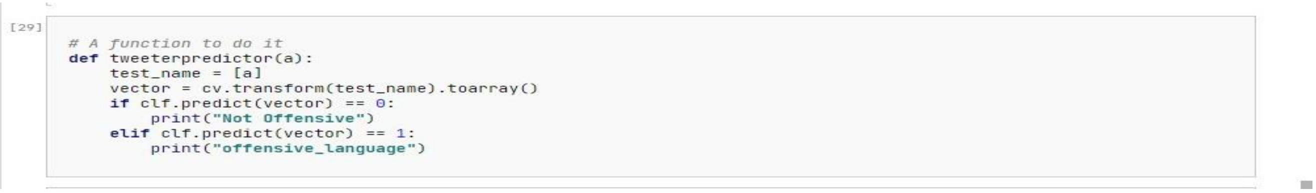
For classification, the logistic regression method is employed to address the problem. When something is likely to happen, such as whether or not an event will take place, it is referred to as the probability of it occurring (Yes or No). When applied to the proposed research, it can, for example, provide binary outcomes, such as the ability to forecast whether a student would be granted a placement or not, or in our case to detect whether the given sentence is offensive of not. Which is a binomial classification issue? It also has the capacity to predict the result of ordinal events. The fact that logistic regression is used to predict categorical values, which is the most common kind of prediction, is the most significant characteristic of the method. Let's start using logistic regression as a starting point for our analysis. Note that this algorithm is not for text classification but as a starting point it is simple to start to learn basic flow how things work in Machine learning.

In this case, logistic regression will be used to train our model:



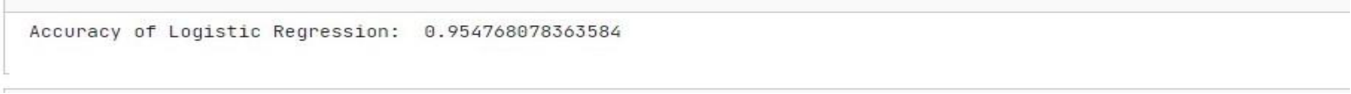
```
[27]
# Logistic Regression Classifier
from sklearn.linear_model import LogisticRegression
clf = LogisticRegression(random_state = 0)
clf.fit(X_train,y_train)

LogisticRegression(random_state=0)
```

*Figure 11 Fitting Logistic Regression*



```
[29]
# A function to do it
def tweeterpredictor(a):
    test_name = [a]
    vector = cv.transform(test_name).toarray()
    if clf.predict(vector) == 0:
        print("Not Offensive")
    elif clf.predict(vector) == 1:
        print("offensive_language")
```

***Figure 10 Prediction Using Logistic Regression***



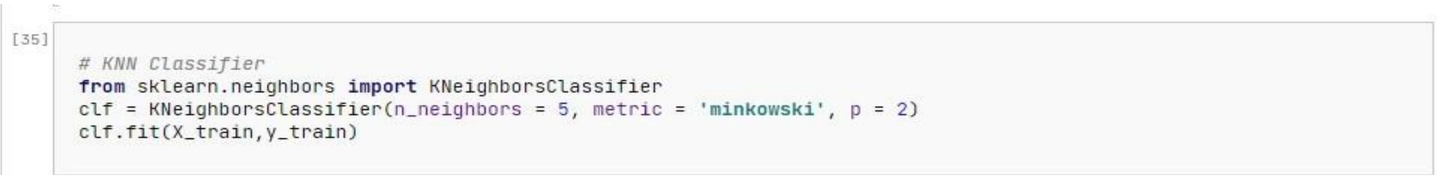
```
Accuracy of Logistic Regression:  0.954768078363584
```

*Figure 12 Logistic Regression Accuracy*

Logistic regression give us best accuracy but it’s not good for text classification as we mention earlier we used Logistic regression for starting ML journey .

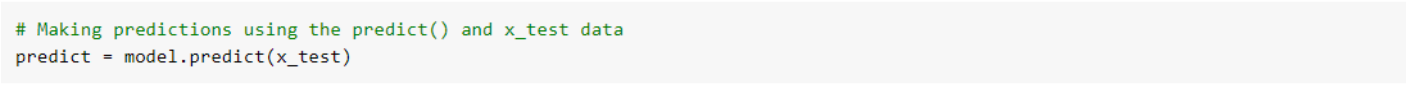
### 3.2.2 K – Nearest Neighbor

Any of the dataset's distributions are completely outside of the control and responsibility of KNN (Cheng et al., 2014). As soon as new data points are collected, they are assigned to a group known as data points, with the classification issue serving as the basis for this assignment. Making a KNN model is a simple and low-cost process that takes little time and effort. It has the capability of dealing with large datasets. In order to determine the distance, it makes use of the k value. When it comes to finding the k value, there is no secret formula to follow.



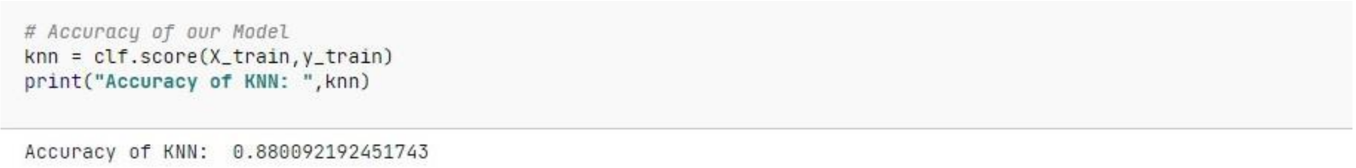
```
[35]
# KNN Classifier
from sklearn.neighbors import KNeighborsClassifier
clf = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
clf.fit(X_train,y_train)
```

*Figure 13 Fitting KNN model*



```
# Making predictions using the predict() and x_test data
predict = model.predict(x_test)
```

*Figure 14 Predection for KNN*



```
# Accuracy of our Model
knn = clf.score(X_train,y_train)
print("Accuracy of KNN: ",knn)
```

```
Accuracy of KNN:  0.880092192451743
```

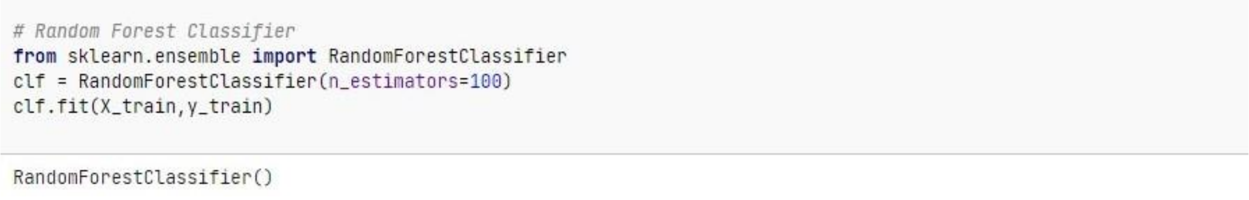
*Figure 15 KNN Accuracy*

The Accuracy of KNN is good but further we will improve it to try other Machine learning algorithm.

### 3.2.3 Support Victor Machine

### 3.2.4 Random Forest

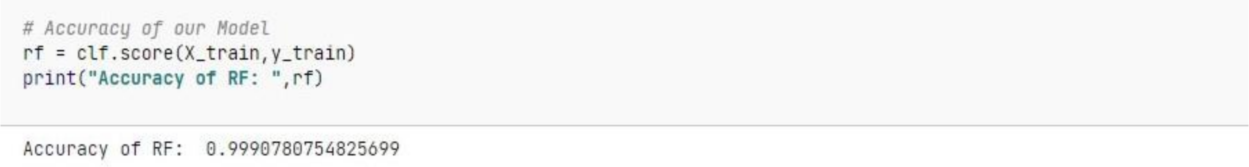
For classification and regression issues, for example, an ensemble learning strategy known as random forest (also known as random choice forest) is used to train a large number of decision trees (Biau & Scornet, 2016). Random forests are used to tackle classification, regression, and other problems. Random forest classification provides the class that has the greatest number of trees when it is utilized for classification.



```
# Random Forest Classifier
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier(n_estimators=100)
clf.fit(X_train,y_train)
```

```
RandomForestClassifier()
```

***Figure 16 Fitting Random Forest***



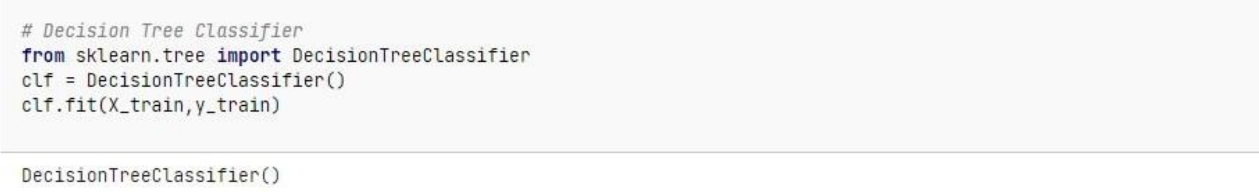
```
# Accuracy of our Model
rf = clf.score(X_train,y_train)
print("Accuracy of RF: ",rf)
```

```
Accuracy of RF:  0.9990780754825699
```

***Figure 17 Random Forest Accuracy***

### 3.2.5 Decision Tree

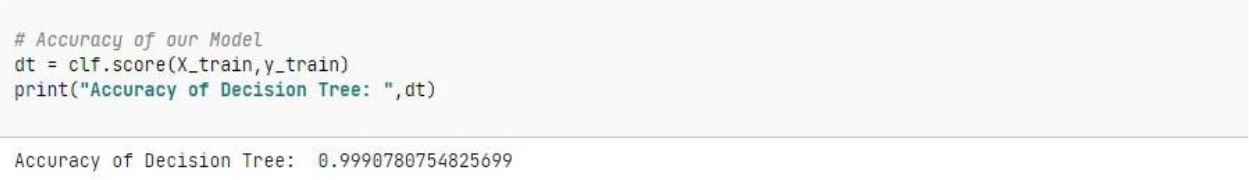
Using a tree-like model of actions and their potential consequences, such as chance event outcomes, resource costs, and utility, a decision tree can be used to aid in decision-making processes. It is one method of demonstrating an algorithm that is solely composed of conditional control statements.



```
# Decision Tree Classifier
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier()
clf.fit(X_train,y_train)
```

```
DecisionTreeClassifier()
```

*Figure 18 Fitting decision Tree*



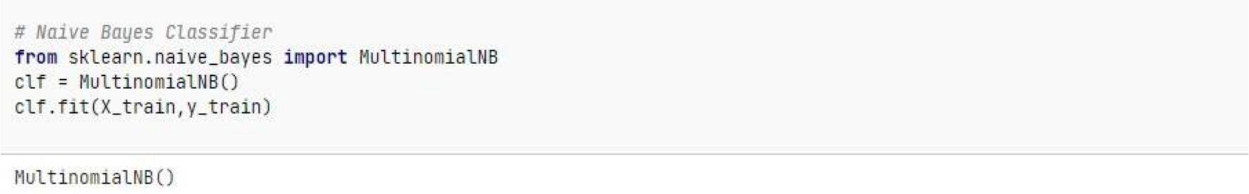
```
# Accuracy of our Model
dt = clf.score(X_train,y_train)
print("Accuracy of Decision Tree: ",dt)
```

```
Accuracy of Decision Tree:  0.9990780754825699
```

*Figure 19 Decision Tree accuracy*

### 3.2.6 Naïve byes

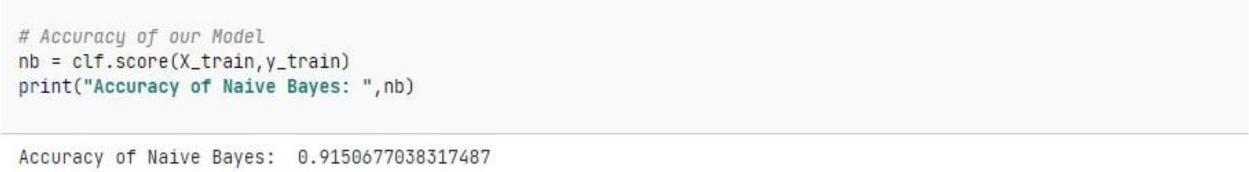
The Naive Bayes approach is a conditional probability-based strategy that is easy to implement. It makes use of a probability table that is updated on a regular basis in response to training data. The Naive Bayes algorithm is a straightforward technique that yields the best possible results in the lowest amount of time possible. It has no effect on them since the Naive Bayes approach does not take into account irrelevant qualities. In addition, we will utilize Naive Bayes to forecast the occurrence of a state in the planned research project.



```
# Naive Bayes Classifier
from sklearn.naive_bayes import MultinomialNB
clf = MultinomialNB()
clf.fit(X_train,y_train)
```

```
MultinomialNB()
```

*Figure 20 Fitting Naive Bayes*



```
# Accuracy of our Model
nb = clf.score(X_train,y_train)
print("Accuracy of Naive Bayes: ",nb)
```

```
Accuracy of Naive Bayes:  0.9150677038317487
```

***Figure 21 Naive Bayes Accuracy***

## 3.3 **Proposed Solution**

## 3.4 **Diagram of Proposed Solution**

# CHAPTER 4: Design & Analysis

In chapter number four, Data set is the main part when you want to tarin ML model so first we will describe data set involve in our work, then we will show the result that we have achieved, then we will do comparative analysis of previous work and the one we did. Finally, we will describe one of the most important methods for evaluation of the model accuracy, precision, recall and f1-score so mainly we will focus on confusion matrix the best method to evaluate ML model.

4.1 **Data Set Description**

4.2 **Results**

4.3 **Comparative Analysis**

4.4 **Evaluation Methods**

# CHAPTER 5: Implementations

In this section of the document, we will discuss about data preprocessing, also highlight the concept of train data and test data respectively and give brief intro to some of the main NLP techniques/methods. Like word embedding, big of words, tokenization, vectorization, and POS etc. I the implementations section we will also describe project specification, system requirements. Finally, we will attach project snap shoot so that will provide clear idea that what we did in our project.

## 5.1 Data Preprocessing

Data preprocessing is the second and most important step of Machine leaning. The more cleanly your text the more accurately machine will understand and the more accurate result it will generate. In this step our main objective is to clean text with any aspect like removing null values, duplicate, special character and many more. So let's understand the whole process by describing it in more detailed.

As our data set has to features the first one is the "tweet" and the second one is the class/label which is offensive or not offensive. Thus we need to pre-process the tweet column for training and the testing set. If we fed clean text to our algorithm then it will generate more reliable and best result as compared to fed dummy or garbage data where our model will not be able to interpret and gather much information about the data as a result it will produce bad and false result.

Below are the main steps involve in text pre-processing that can be considered to provide better version of the twitter text.

- **Tokenization:**
  It is the process of splitting the whole sentences or paragraph or sequence of string into respective tokens which represent the individual words. It is very important step it gives us total amount of words and also frequency of the words how many times specific words comes in. we can also extract words by this method tokenization.
- **Stemming:**
  It is the process of converting or reducing the inflected and derivationally forms of related verbs. It may add affixes to the base form of the word or removing the suffix from a word.
- Lemmatization:
- Lover Case Conversion:
- Stop Words Removal:
- Punctuation Removal:

- User and URL Removal:
- Special Characters Removal:
- Numbers and Emoji Handling:
- White Space Removal:

## 5.2 Train test split

## 5.3 Feature Extraction

## 5.4 Model Building

## 5.5 Project specifications

The following are the project specification that we used in our project.

### 5.5.1 Training and testing data set

Offensive language identification dataset which consist of 14.200 English tweets created in 2019.

### 5.5.2 Python 3

Latest version of Python.

Anaconda distribution specifically Jupyter Note book.

ML and DL based model to detect offensive language.

- SVM
- KNN
- Logistic regression
- Decision tree
- Naïve Bayes

### 5.5.3 Python libraries

- NumPy
- Pandas

- Scikit-learn
- Karas
- Streamlet
- TkInter

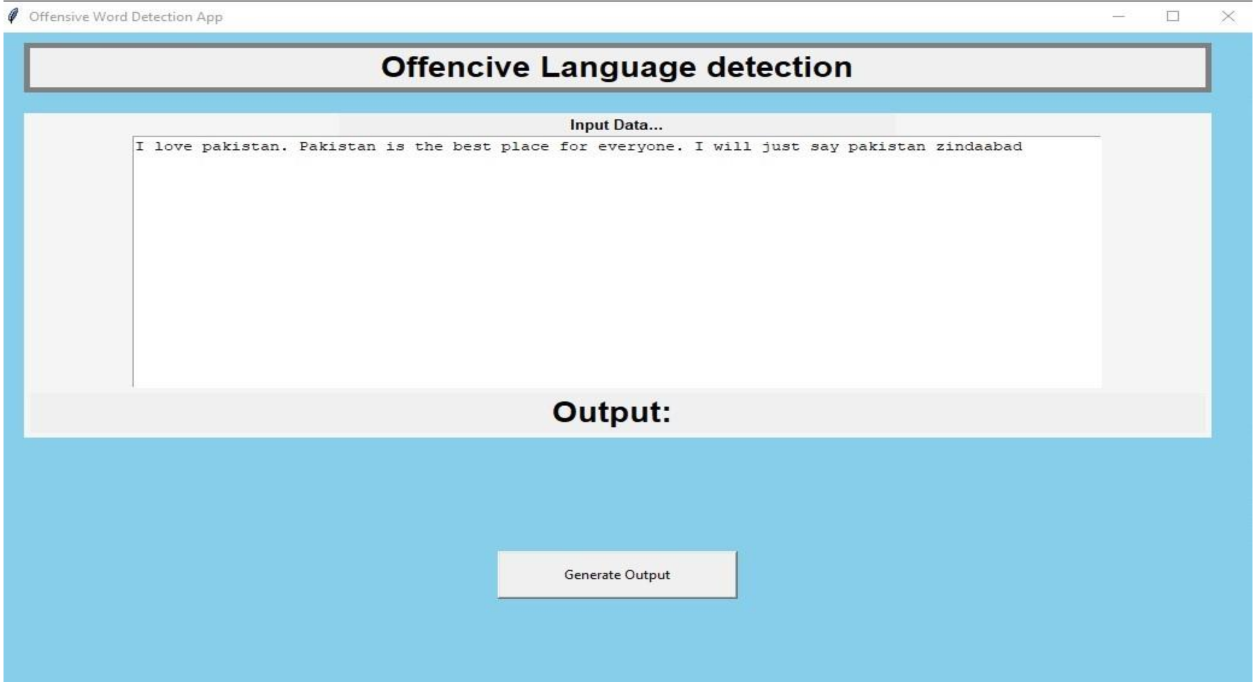
### 5.5.4 Evaluation

- Confusion matrix
- F1-score

## 5.6 Project Snap Shoots

### 5.6.1 Version 1.0.0

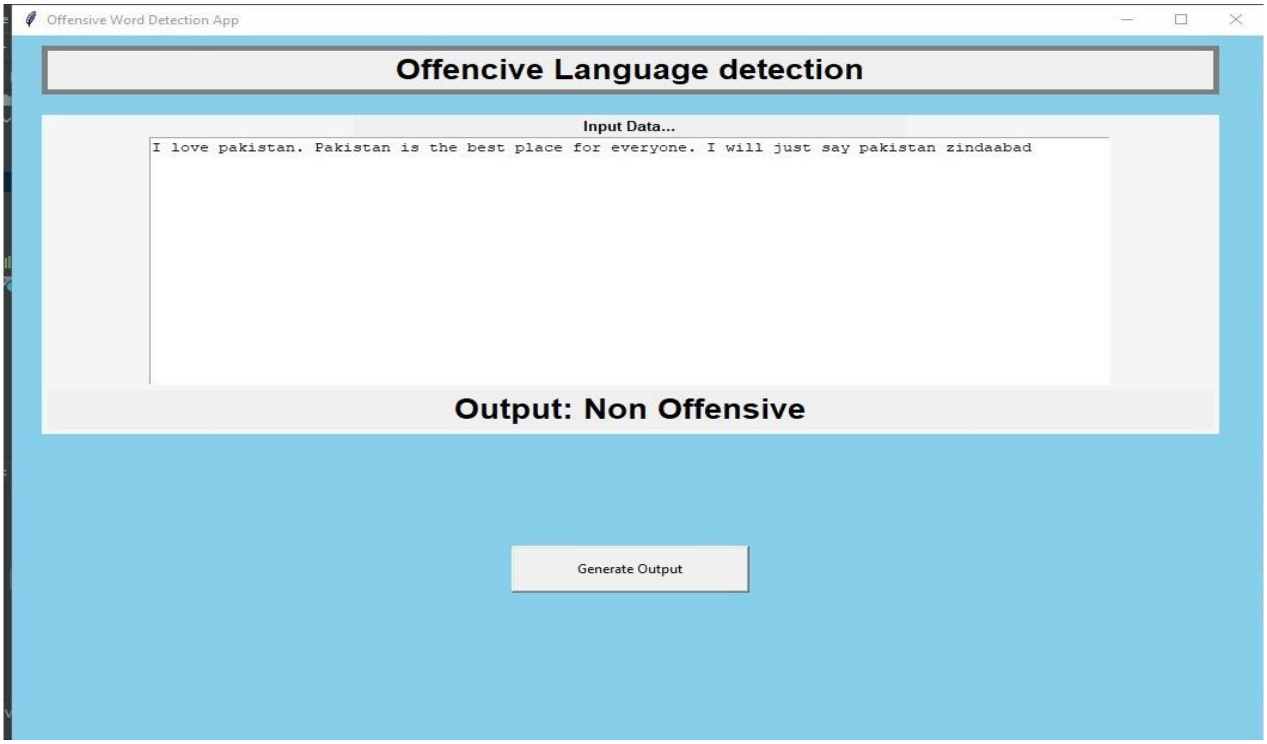
In this version we used Tkinter for GUI  to show the result of our model how correctly it identify the offensive language.



I love pakistan. Pakistan is the best place for everyone. I will just say pakistan zindaabad

Output:

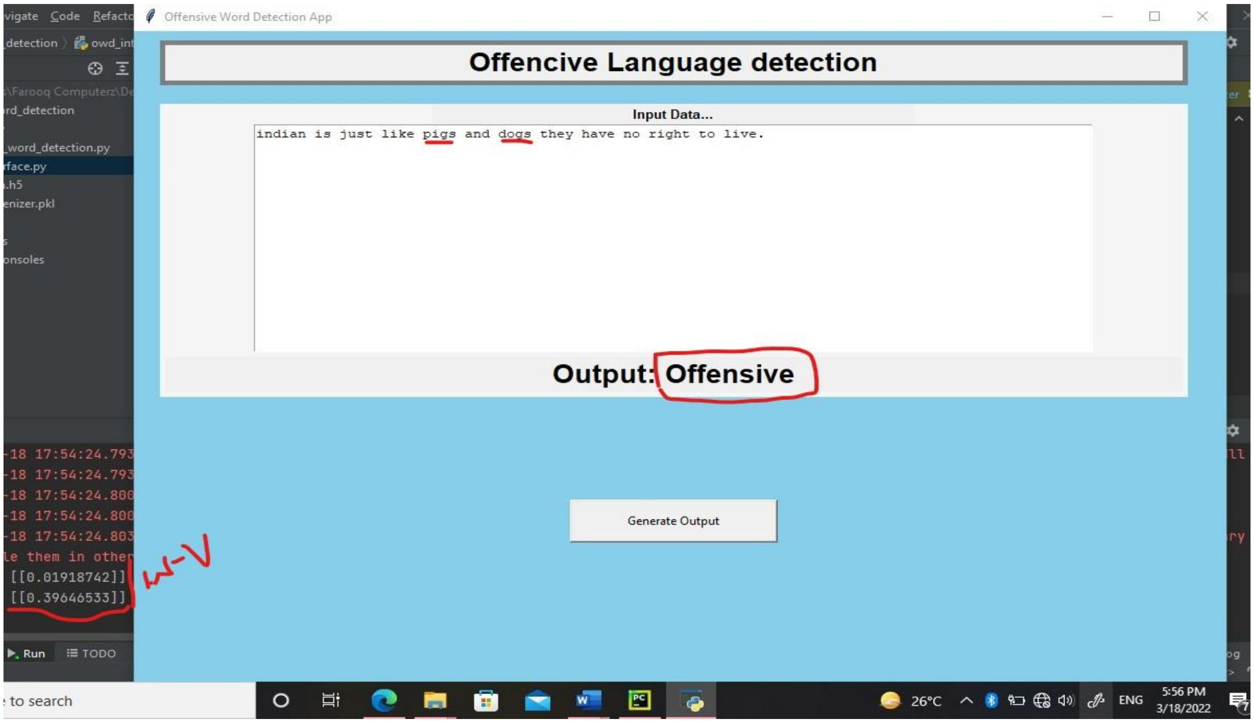
*Figure 22 Output  window*



I love pakistan. Pakistan is the best place for everyone. I will just say pakistan zindaabad

Output: Non Offensive

*Figure 23 Output with window showing Non offensive result*



Offencive Language detection

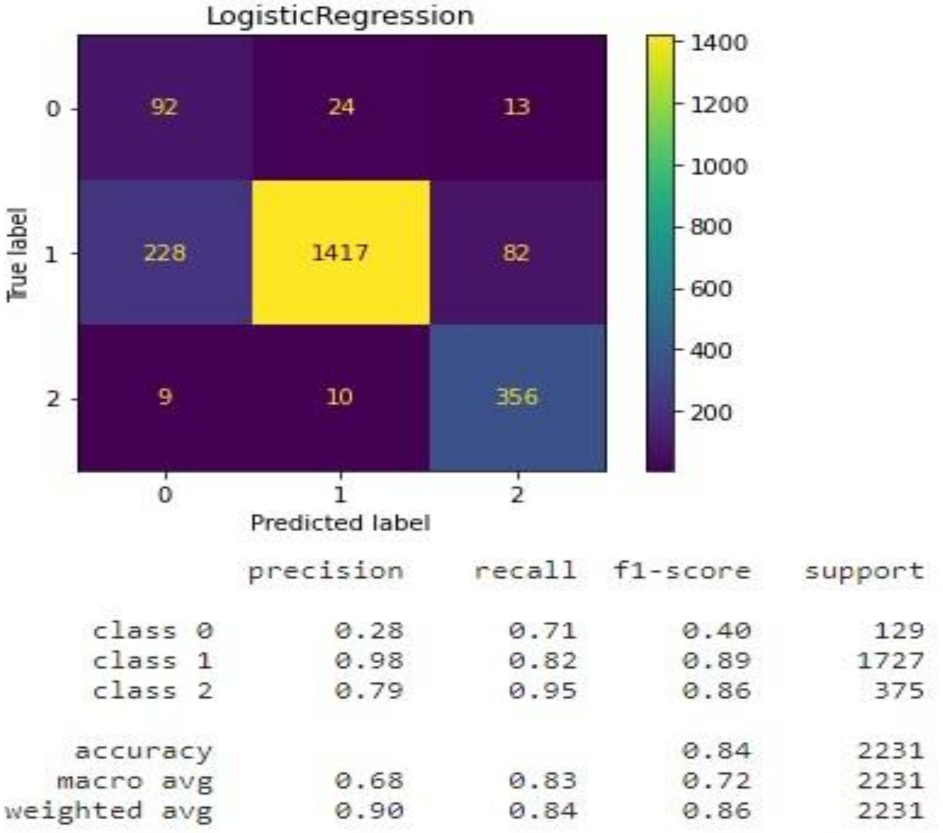
Input Data...

indian is just like pigs and dogs they have no right to live.

Output: Offensive

```
-18 17:54:24.793
-18 17:54:24.793
-18 17:54:24.800
-18 17:54:24.800
-18 17:54:24.803
le them in other
 [[0.01918742]]
 [[0.39646533]]
```

*Figure 24 Output window showing offensive result*

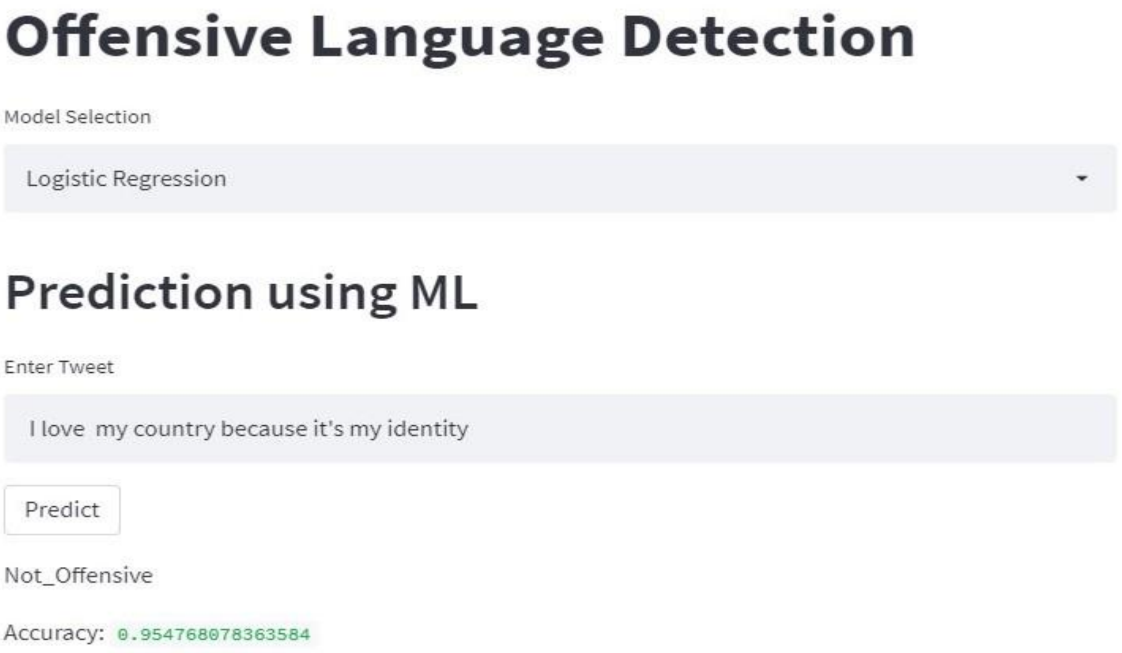


```
              precision    recall  f1-score   support

     class 0       0.28      0.71      0.40       129
     class 1       0.98      0.82      0.89      1727
     class 2       0.79      0.95      0.86       375

    accuracy                           0.84      2231
   macro avg       0.68      0.83      0.72      2231
weighted avg       0.90      0.84      0.86      2231
```

*Figure 25 Confusion matrix for LR*

### 5.6.2 Version 1.1.0

In version 1.1.0 we use streamlit to show the over all performance of our models five models



# Offensive Language Detection

Model Selection

Logistic Regression

## Prediction using ML

Enter Tweet

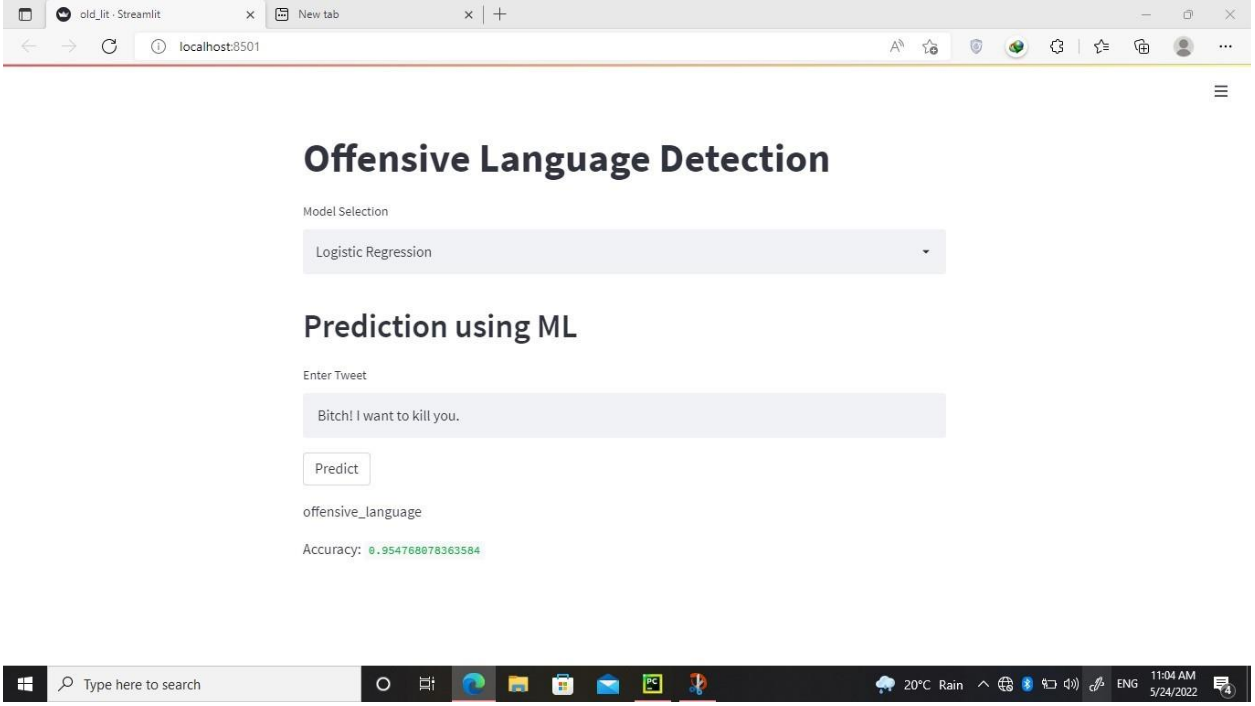
I love  my country because it's my identity

Predict

Not_Offensive

Accuracy: `0.954768078363584`

***Figure 26 Logistic Regression Not-Offensive***



# Offensive Language Detection

Model Selection

Logistic Regression

## Prediction using ML

Enter Tweet

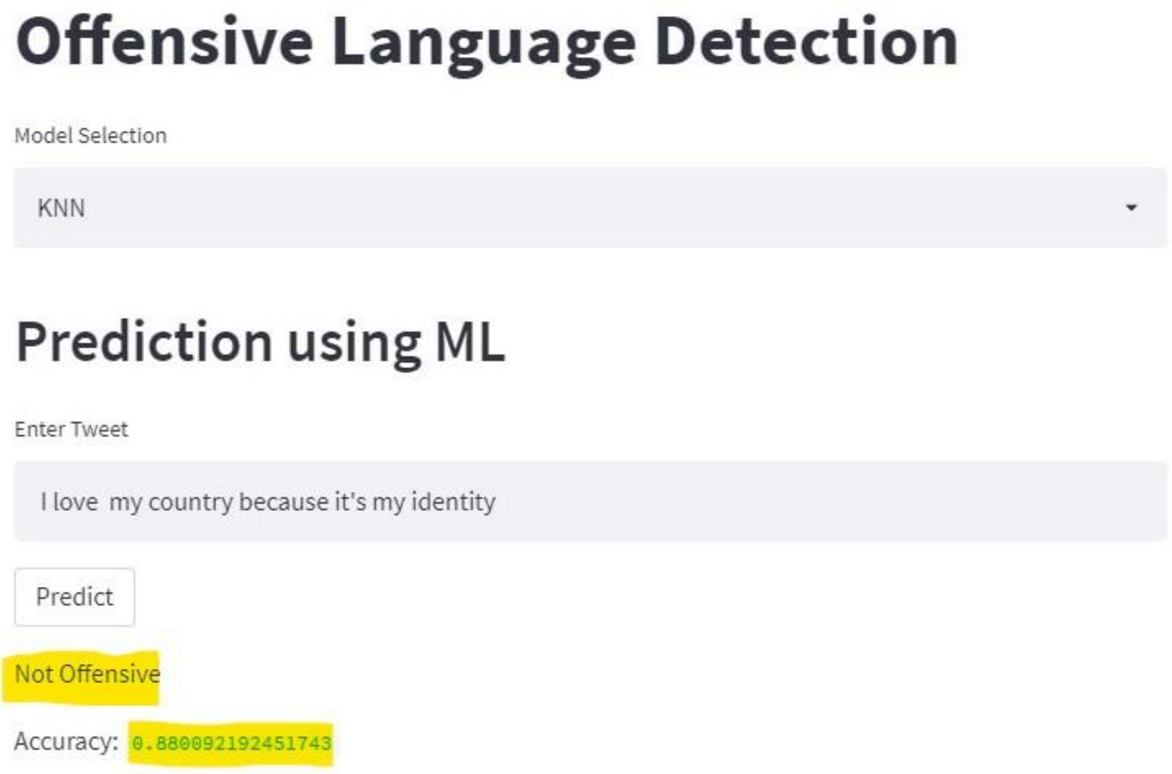
Bitch! I want to kill you.

Predict

offensive_language

Accuracy: `0.954768078363584`

***Figure 27 Logistic Regression Offensive***



Offensive Language Detection

Model Selection

KNN

Prediction using ML

Enter Tweet

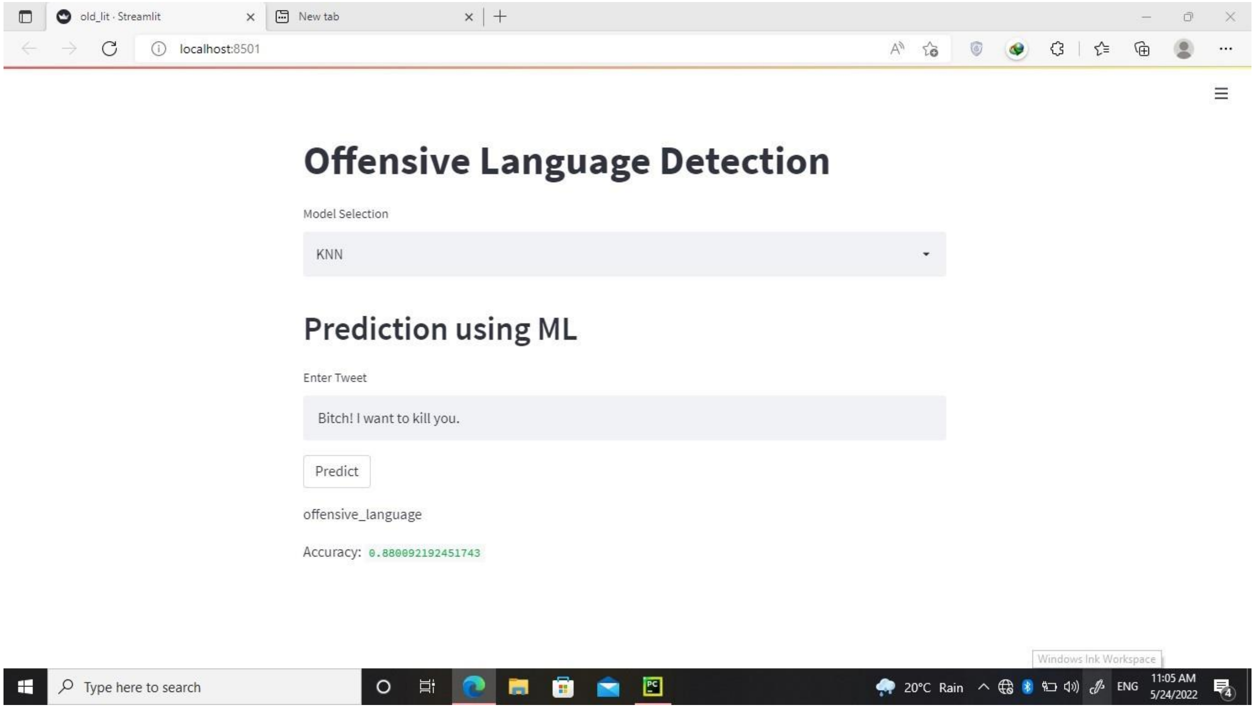
I love  my country because it's my identity

Predict

Not Offensive

Accuracy: 0.880092192451743

*Figure 28 KNN Not-Offensive*



Offensive Language Detection

Model Selection

KNN

Prediction using ML

Enter Tweet

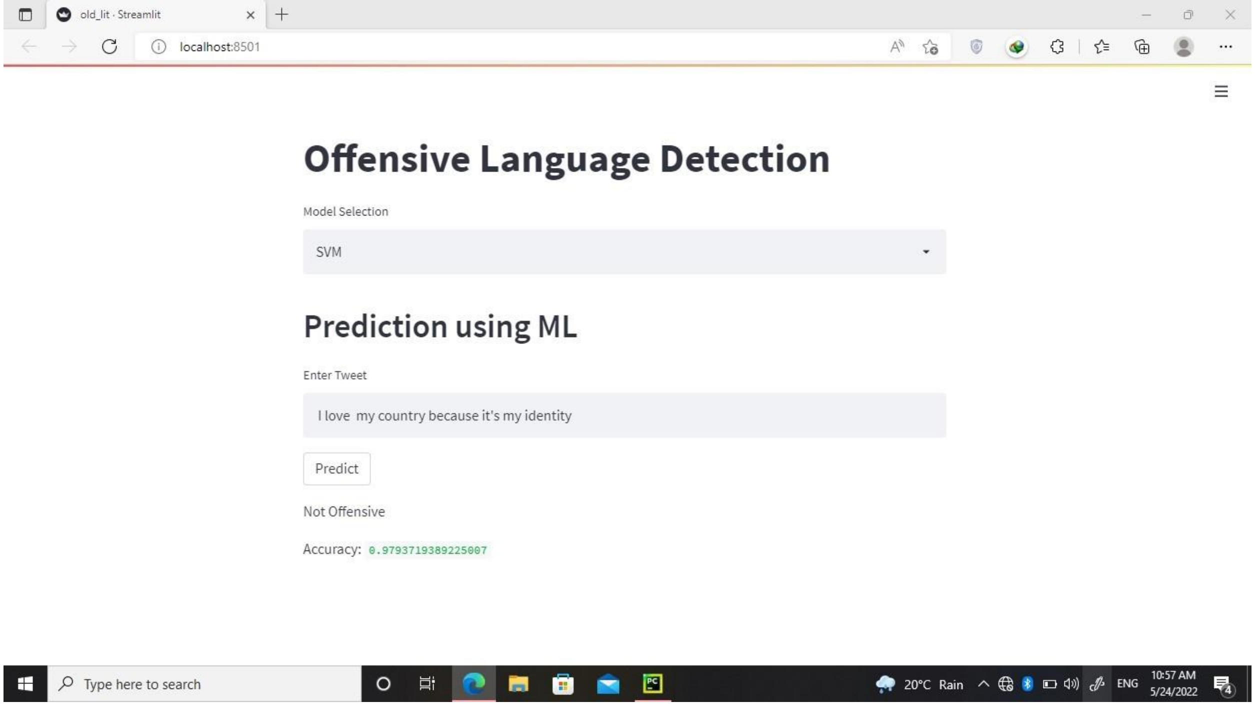
Bitch! I want to kill you.

Predict

offensive_language

Accuracy: 0.880092192451743

*Figure 29 KNN Offensive*



# Offensive Language Detection

Model Selection

SVM

## Prediction using ML

Enter Tweet

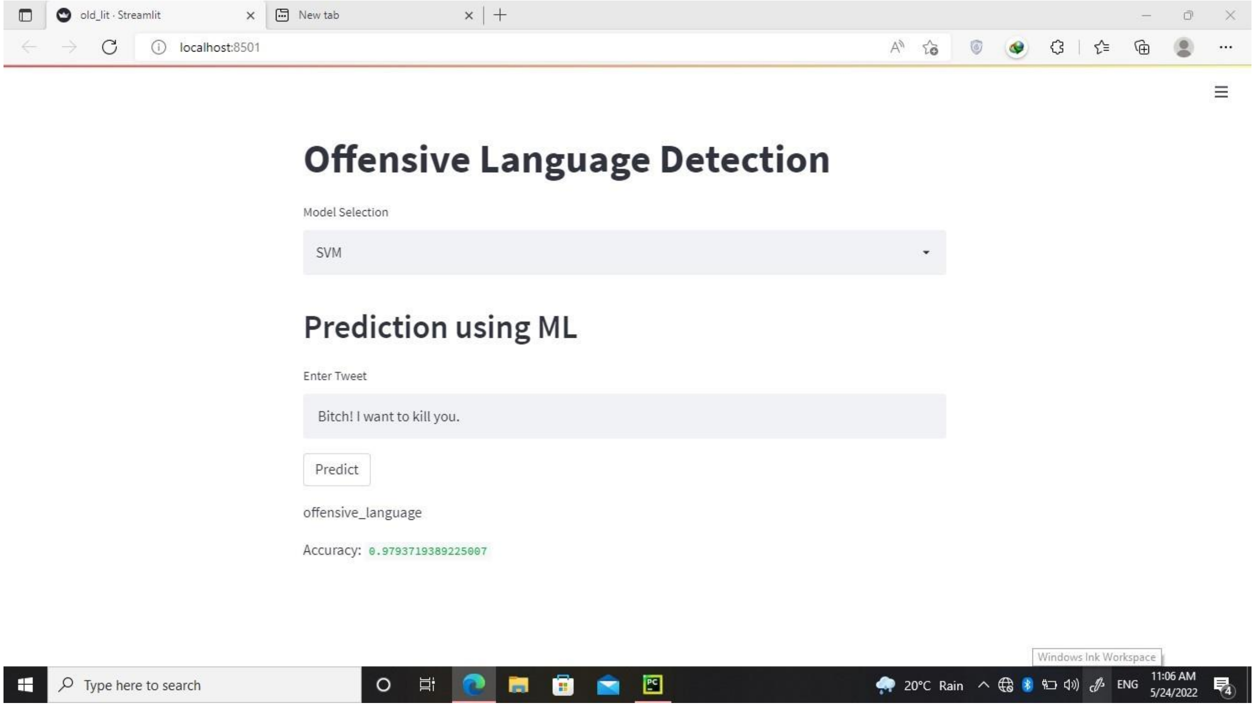
I love  my country because it's my identity

Predict

Not Offensive

Accuracy: `0.9793719389225007`

*Figure 30 SVM Not-Offensive*



# Offensive Language Detection

Model Selection

SVM

## Prediction using ML

Enter Tweet

Bitch! I want to kill you.

Predict

offensive_language

Accuracy: `0.9793719389225007`

*Figure 31 SVM Offensive*

All others model is the same output we show few of them for sample to give celerity about our work. It show s accuracy and  also say either your statement is offensive or not.

# REFERENCES

[1] Nockleby JT. Hate Speech. Encyclopedia of the American Constitution 2000; 3 1277-79. Google scholar.

[2] Community standard:https:www.facebook.com/communitystandards/objectionable.content.

[3] de Gibert O,perez N, Garcia-Pablos A, Cuadros M, Hate Speech dataset from a white supremacy forum. In 2nd workshop on abusive language online @EMNLP; 2018.

[4] Ultimate guide to deal with text data using python for data scientists and engineers available it: www.analyticsvidhya.com

[5] K., Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in 5th international AAAI conference on web logs and social media, 2011.

[6] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in Proceedings of the 2012 conference of the North American chappter of the association for computational linguistics: Human language technologies. Association for Computational Linguistics, 2012, pp. 656–666.

[7] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in European Conference on Information Retrieval. Springer, 2013, pp. 693–696.

[8] P. Burnap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," Policy & Internet, vol. 7, no. 2, pp. 223–242, 2015.

[9] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in Proceedings of the 24th international conference on world wide web, 2015, pp. 29–30.

[10] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in Proceedings of the 25th international conference on world wide web, 2016, pp. 145–153.

[11] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Eleventh international aaai conference on web and social media, 2017

[12] S. Malmasi and M. Zampieri, "Detecting hate speech in social media," arXiv preprint arXiv: 1712.06427, 2017.

# PLAGIARISM REPORT BY TURNIT