A Project Report on

TWITER USER BULLYING DETECTION

Submitted in partial fulfilment of the requirement for the award of the degree in

MASTER OF COMPUTER APPLICATION

Department of Information Technology
and
Computer Applications

Submitted To:	Submitted By:
Mr. Rahul Ranjan	**Vijay Pratap Singh**
(Assistant Professor)	21MCAN055

DECLARATION

We hereby declare that the work described in this project work, entitled “**TWITTER USER BULLYING DETECTION"** which is submitted by us in partial fulfilment for the award of MASTER OF COMPUTER APPLICATION TO THE JECRC UNIVERSITY JAIPUR, IS the result of work done by us under the guidance of **Mr. Rahul Ranjan**, Assistant Professor.

The work is original and has not been submitted for any Degree of this or any other university.

Candidate Signature: Vijay Pratap Singh

Branch: IT **Student ID:** 21MCAN055

Submitted To:
Department of Information Technology, JECRC University, Jaipur
State: Rajasthan

ABSTRACT

With growth in Internet Technology, social media has highly gained popularity as a medium of interaction. With increase in social media conversation, there is increase in activities such as aggressive and intentional actions performed via digital communication such as sending abusive messages and posting immoral comments. User Base for social media such as Facebook, twitter has increased multiple times in recent past. Also, there is a high increase in unstructured data in the means of reviews and comments in online portals. With these huge data growing there is a lot of scope to use big data technologies to analyses these data. In our project, we propose to handle unstructured data using machine learning algorithms. We use these technologies to address cyberbullying on Twitter. In the proposed system, random forest regressor, a machine learning algorithm is used to train the dataset consisting of a collection of abusive comments. By accurately predicting the abusive comments used the system can be used by social media platforms to block such abusive comments. Thus, improving the user experience for using and accessing social media platforms.

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

CNN	Convolutional Neural
NetworkML	Machine Learning
ReLU	Rectified Linear Unit
SARSA	State-Action-Reward-State-
Action DDPG	Deep Deterministic Policy
Gradient	

TABLE OF CONTENTS

CHAPTER-1
INTRODUCTION

1.1 GENERAL

Social media has made it very easy for us to communicate quickly and easily with family, friends and acquaintances, as well as sharing experiences and letting others know of our opinions and beliefs. These opinions and beliefs may be about world events or local affairs, politics or religion, interests, affiliations, organizations, products, people and a wide variety of other topics. Our conversations and comments can be closely targeted or widely broadcast to the point that depending on the subject, they can go viral. Unfortunately, social media is also widely used by abusers, for exactly the reasons listed above. Many perpetrators 'hide' behind the fact that they may not be able to be readily identified, saying things that they wouldn't consider saying face-to-face, which could be regarded as cowardly.

Online abuse takes several forms, and victims are not confined to public figures. They can do any job, be of any age, gender, sexual orientation or social or ethnic background, and live anywhere.

Cyberbullying can occur online only, or as part of more general bullying. Cyberbullies may be people who are known to you or anonymous. Like all bullies, they frequency try to persuade others to join in. You could be bullied for your religious or political beliefs, race or skin colour, body image, if you have a mental or physical disability or for no apparent reason whatsoever.

Cyberbullying generally comprises sending threatening or otherwise nasty messages or other communications to people via social media, gaming sites, text or email, posting embarrassing or humiliating video on hosting sites such as YouTube or Vimeo, or harassing through repeated texts, instant messages or chats. Increasingly, it is perpetrated by posting or forwarding images, video or private details obtained via sexting, without the victim's permission. Some cyberbullies set up Facebook pages and other social media accounts purelyto bully others.

1.2 TECHNOLOGY USED

1.2.1 Machine Learning

Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications,such as in medicine, email filtering, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

Machine learning plays an important role in cybersecurity and online fraud detection. Because of growing monetary online frauds, companies like PayPal have started using machine learning techniques for protection against the money laundering. The prediction problem of the model for fraud detection can be divided into two types: classification and regression. Some of the most used machine learning approaches for this type of prediction problems are Logistic Regression, Decision Tree, Random Forest Tree, and Neural Networks.

Modern day machine learning has two objectives, one is to classify data based on models which have been developed, the other purpose is to make predictions for future outcomes based on these models. A hypothetical algorithm specific to classifying data may use computer vision of moles coupled with supervised learning in order to train it to classify the cancerous moles. Where as machine learning algorithm for stock trading may inform the trader of future potential predictions.

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: The ability to learn. Machine learning is actively beingused today, perhaps in many more places than one would expect.

A subset of machine learning is closely related to computational statistics, whichfocuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve theproblem at hand; on the computer's part, no learning is needed.

The machine learning field is continuously evolving. And along with evolution comes a rise in demand and importance. There is one crucial reason why data scientists need machine learning, and that is: ‘High-value predictions that can guide better decisions and smart actionsin real-time without human intervention.

Machine learning as technology helps analyze large chunks of data, easing the tasks of data scientists in an automated process and is gaining a lot of prominence and recognition. Machine learning has changed the way data extraction and interpretation works by involving automatic sets of generic methods that have replaced traditional statistical techniques.

For more advanced tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than having human programmers specify every needed step.

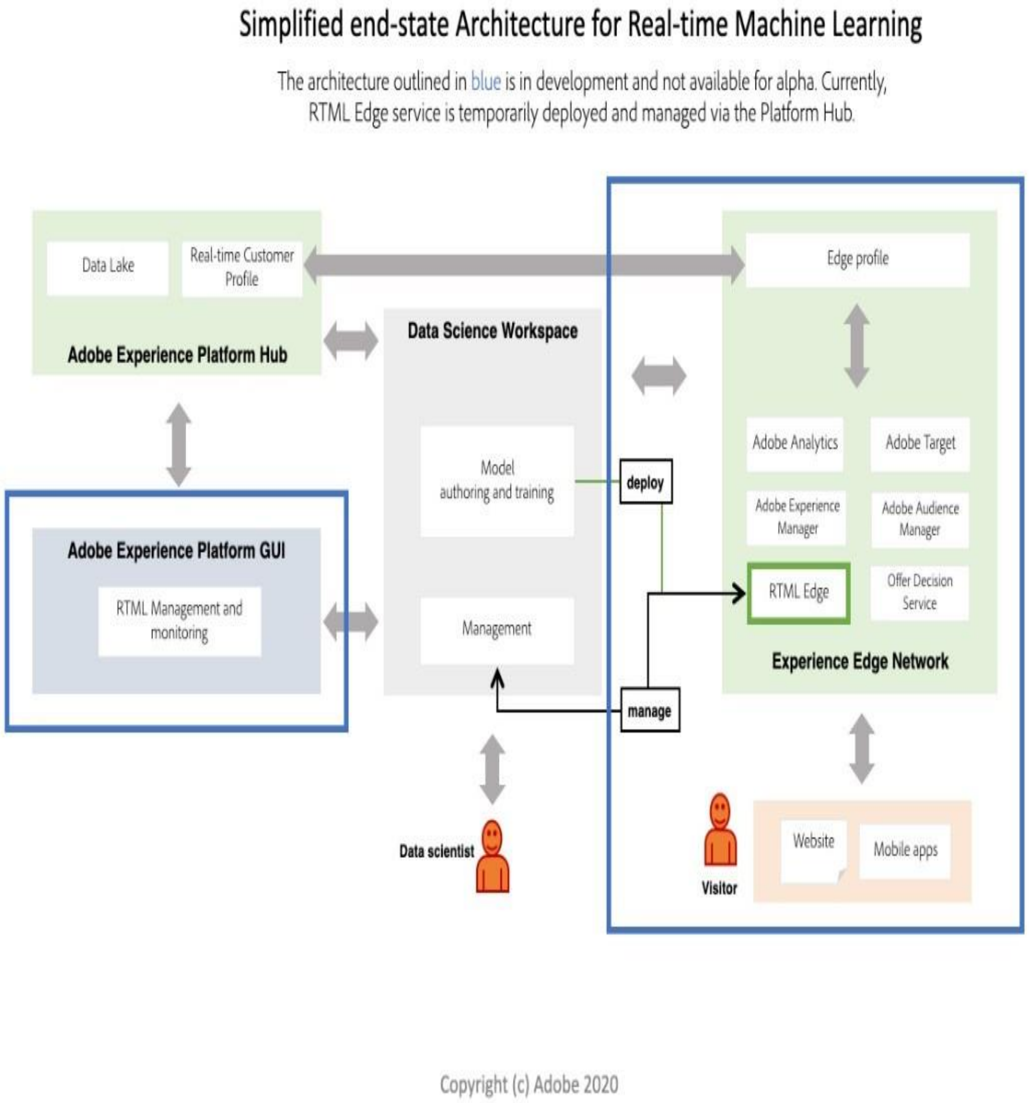

Simplified end-state Architecture for Real-time Machine Learning
The architecture outlined in blue is in development and not available for alpha. Currently, RTML Edge service is temporarily deployed and managed via the Platform Hub.
Data Lake
Real-time Customer Profile
Adobe Experience Platform Hub
Data Science Workspace
Model authoring and training
Management
Adobe Experience Platform GUI
RTML Management and monitoring
deploy
manage
Edge profile
Adobe Analytics
Adobe Target
Adobe Experience Manager
Adobe Audience Manager
RTML Edge
Offer Decision Service
Experience Edge Network
Data scientist
Visitor
Website
Mobile apps
Copyright (c) Adobe 2020

Figure 1.1 Machine learning architecture

1.2.1.1 Machine Learning approaches

Machine learning approaches are traditionally divided into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system:

Supervised learning

Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs. The data is known as training data, and consists of a set of training examples. Each training example has one or more inputs and the desired output, also known as a supervisory signal. In the mathematical model, each training example is represented by an array or vector, sometimes called a feature vector, and the training data is represented by a matrix. Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict theoutput associated with new inputs. An optimal function will allow the algorithm to correctly determine the output for inputs that were not a part of the training data. An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task. Machine Learning algorithms, classification and regression. Classification algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a range. As an example, for a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email.

Similarity Learning is an area of supervised machine learning closely related to regression and classification, but the goal is to learn from examples using a similarity function that measures how similar or related two objects are. It has applications in ranking, recommendation systems, visual identity tracking, face verification, and speaker verification.

Supervised learning as the name indicates the presence of a supervisor as a teacher. Basically, supervised learning is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcomefrom labelled data.

Supervised learning is where there are input variables (x) and an output variable (Y)and an algorithm is used to learn the mapping function from the input to the output.

Y = f(X)

The goal is to approximate the mapping function so well that when there is a new inputdata (x) that the output variables (Y) for that data can be predicted easily.

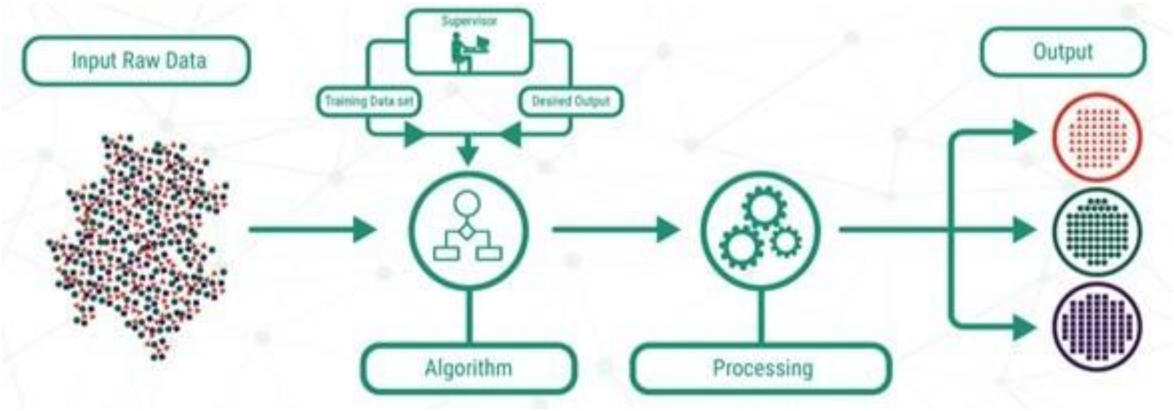


Figure 1.2 Supervised learning flowchart

Unsupervised learning

Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. The algorithms, therefore, learn from test data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data. A central application of unsupervised learning is in the field of density estimation in statistics, such as finding the probability density function. Though unsupervised learning encompasses otherdomains involving summarizing and explaining data features.

Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data. Unlike supervised learning, no teacher isprovided that means no training will be given to the machine. Therefore, machine is restrictedto find the hidden structure in unlabeled data by itself.

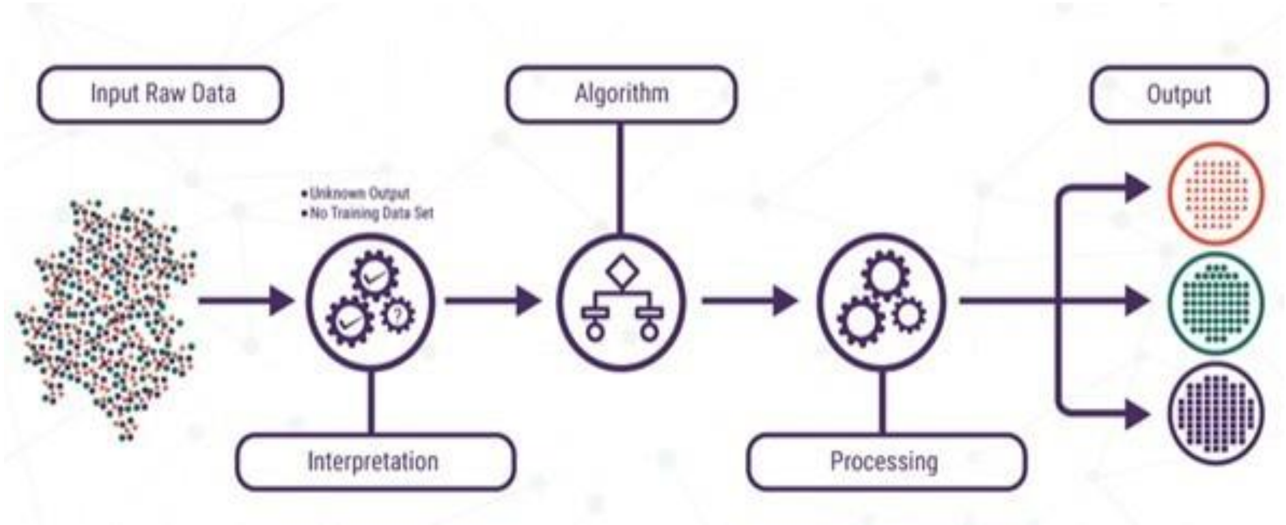


Figure1.3 Unsupervised learning flowchart.

Semi-supervised Learning

Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Some of the training examples are missing training labels, yet many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce a considerable improvement in learning accuracy.

Reinforcement learning

Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward. Due to its generality, the field is studied in many other disciplines, such as game theory, control theory, operations research, information theory, simulation-based optimization, multi-agent systems, swarm intelligence, statistics and genetic algorithms. In machine learning, the environment is typically represented as a Markov decision process (MDP).Many reinforcement learning algorithms use dynamic programming techniques. Reinforcement learning algorithms do not assume knowledge of an exact mathematical model of the MDP, and are used when exact models are infeasible. Reinforcement learning algorithms are used in autonomous vehicles or in learning to play a game against a human opponent.

A computer program interacts with a dynamic environment in which it must perform a certain goal. As it navigates its problem space, the program is provided feedback that's analogous to rewards, which it tries to maximize.

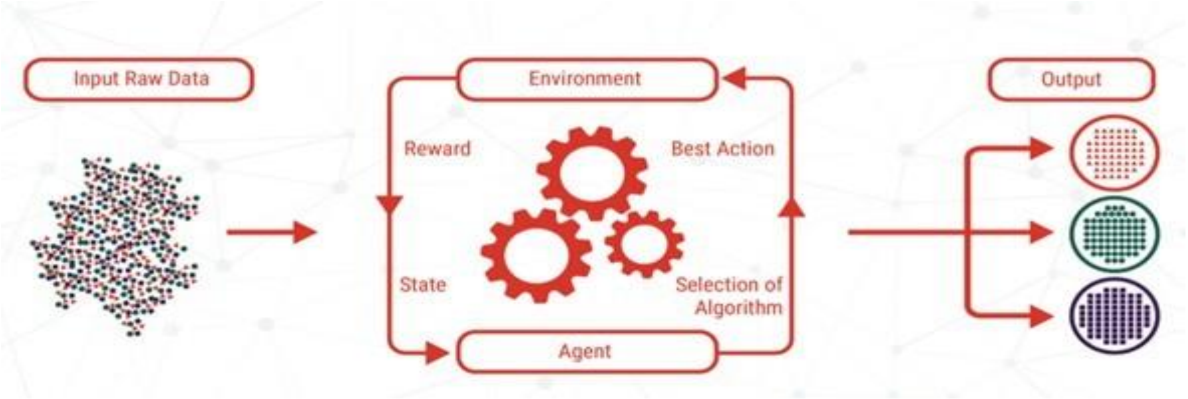


Figure 1.4 Reinforcement learning flowchart

Feature Learning

Several learning algorithms aim at discovering better representations of the inputs provided during training.[47] Classic examples include principal components analysis and cluster analysis. Feature learning algorithms, also called representation learning algorithms, often attempt to preserve the information in their input but also transform it in a way that makes it useful, often as a pre-processing step before performing classification or predictions. This technique allows reconstruction of the inputs coming from the unknown data-generating distribution, while not being necessarily faithful to configurations that are implausible under that distribution. This replaces manual feature engineering, and allows a machine to both learn the features and use them to perform a specific task.

Feature learning can be either supervised or unsupervised. In supervised feature learning, features are learned using labeled input data. Examples include artificial neural networks, multilayer perceptron's, and supervised dictionary learning. In unsupervised feature learning, features are learned with unlabeled input data. Examples include dictionary learning, independent component analysis, autoencoders, matrix factorization[48] and various forms of clustering.

Manifold learning algorithms attempt to do so under the constraint that the learned representation is low-dimensional. Sparse coding algorithms attempt to do so under the constraint that the learned representation is sparse, meaning that the mathematical model has many zeros. Multilinear subspace learning algorithms aim to learn low-dimensional representations directly from tensor representations for multidimensional data, without reshaping them into higher-dimensional vectors. Deep learning algorithms discover multiple levels of representation, or a hierarchy of features, with higher-level, more abstract features defined in terms of (or generating) lower-level features. It has been argued that an intelligent machine is one that learns a representation that disentangles the underlying factors of variation that explain the observed data.

Feature learning is motivated by the fact that machine learning tasks such as classification often require input that is mathematically and computationally convenient to process. However, real-world data such as images, video, and sensory data has not yielded to attempts to algorithmically define specific features. An alternative is to discover such featuresor representations through examination, without relying on explicit algorithms.

1.3 OBJECTIVE

- To effectively develop a system to help social media platforms detect and identify abusivecomments.
- To make use of random forest regressor, a regression machine learning algorithm foraccurate detection of abusive comments.

CHAPTER -2

LITERATURE SURVEY

2.1 INTRODUCTION

The following shows survey on social media abuse detection. The most popular ofthe existing techniques used for power forecasting.

2.2 LITERATURE SURVEY

Title of the Paper	Author Name	Algorithm	Advantages	Disadvantages
A High-Accuracy Wind Power Forecasting Model [2016]	Shengchen Fang; Hsiao-Dong Chiang	Makes use of Gaussian process models	The system is fully automated giving very good accuracy in results.	The system only makes use of a single machine learning algorithms.
Long-Term Retail Energy Forecasting With Consideration of Residential Customer Attrition [2017]	Jingrui Xie SAS Institute, Cary, NC, USA TaoHong; Joshua Stroud	Makes use of regression algorithms for forecasting	The proposed methodology has been implemented and the results are very accurate.	The system only focuses on the forecasting in residential aspects of energy.
Probabilistic Load Forecasting via Quantile Regression Averaging Forecasts[2018]	Bidong Liu; Jakub Nowotarski; Tao Hong; Rafał Weron	Makes use of sister point forecast method machine learning algorithms.	Compared with several benchmark methods, the proposed approach leads to dominantly better	The system does not focus on power forecasting.

			performance as measured by the pinball loss function and the Winkler score.	
An Ensemble Forecasting Method for the Aggregated Load With Subprofiles [2019]	Yi Wang; Qixin Chen; Mingyang Sun; Chongqing Kang; Qing Xia	Makes use of a single machine learning method	An optimal weighted ensemble approach is employed to combine these forecasts and provide the final forecasting result	The system uses only a single machine learning algorithm.
An attention-based unsupervised adversarial model for review spam detection[2020]	Yuan Gao, Maoguo Gong, Senior Member, IEEE, Yu Xie, and A. K. Qin, Senior Member, IEEE	Makes use of unsupervised model	Identifies spam movie reviews	Not very accurate

2.3 MAJOR DISADVANTAGES IN EXISTING SYSTEM

- The existing system focuses on identifying offensive comments, it is very slow andhigh false positives will be observed.
- The system is not very accurate.
- The main limitation of Naive Bayes is the assumption of independent predictor features**.** Naive Bayes implicitly assumes that all the attributes are mutually independent. In real life, it’s almost impossible that we get a set of predictors that are completely independent or one another.
- If a categorical variable has a category in the test dataset, which was not observed intraining dataset, then the model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as Zero Frequency**.**

CHAPTER – 3

SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

With the prevalence of the Internet, online reviews have become a valuable information resource for people. How- ever, the authenticity of online reviews remains a concern, and deceptive reviews have become one of the most urgent network security problems to be solved. Review spams will mislead users into making suboptimal choices and inflict their trust in online reviews. Most existing research manually extracted features and labeled training samples, which are usually complicated and time- consuming.

This paper focuses primarily on a neglected emerging domain - review, and develops a novel unsupervised spam detection model with an attention mechanism. By extracting the statistical features of reviews, it is revealed that users will express their sentiments ondifferent aspects of movies in reviews. An attention mechanism is introduced in the review embedding, and the conditional generative adversarial network is exploited to learn users' review style for different genres of comments. The experimental results demonstrate the superior performance of the proposed approach.

3.2 DISADVANTAGES OF EXISTING SYSTEM

- The existing system focuses on identifying offensive comments, it is very slow andhigh false positives will be observed.
- It assumes that all predictors are independent. Its estimations can be wrong in somecases so we cannot take probability outputs seriously.
- The system is not very accurate.
- Naive Bayes assumes that all predictors (or features) are independent, rarely happening in real life. ...
- This algorithm faces the 'zero-frequency problem' where it assigns zero probability to a categorical variable whose category in the test data set wasn't available in the training dataset.

3.3 PROPOSED SYSTEM

User Base for social media such as Facebook, twitter has increased multiple times in recent past. Also, there is high increase in unstructured data in the means of reviews and comments in online portal. With these huge data growing there is lot of scope to use big data technologies to analyses these data. In our project, we propose to handle unstructured data using machine learning algorithms. We use these technologies to address cyberbullying in Twitter. In the proposed system, random forest regressor, a machine learning algorithm isused to train the dataset consisting of a collection of abusive comments. By accurately predicting the abusive comments used the system can be used by social media platforms to block such abusive comments. Thus, improving the user experience for using and accessing social media platforms.

3.4 ADVANTAGES OF PROPOSED SYSTEM:

- ➢ A technological solution for detecting abusive comments on social media and low falsepositives will be observed.
- ➢ Random forest regressor is used which provides accurate detection.

3.5APPLICATIONS

- Social media platforms.
- Cardiovascular Disease Prediction.
- Diabetes Prediction.
- Breast Cancer Prediction.
- Stock Market Prediction
- Stock Market Sentiment Analysis

CHAPTER-4
SYSTEM DESIGN

DETAILED DESIGN OF THE PROJECT:

This chapter describes the overall and the detailed architectural design. It also describeseach module that is to be implemented along with Data Flow diagram.

4.1 ARCHITECTURE DIAGRAM:

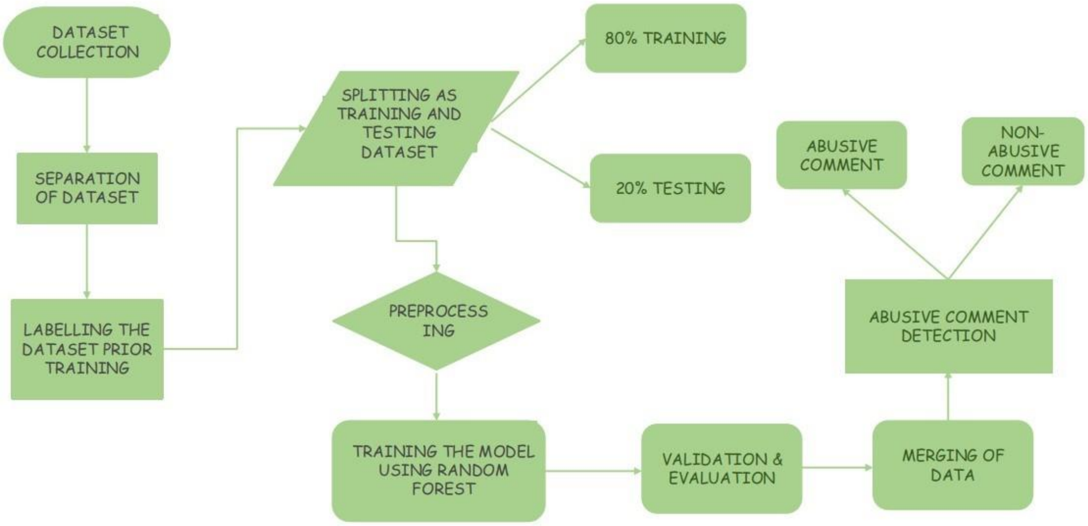


Figure 4.1 System Architecture

4.2 WORKING:

The aim of this project is to investigate and implement algorithms that could possibly detect and identify the abusive comments on social media. Data mining techniques and machine learning algorithms can be used for the prediction and detection of abusive comments. In this project the initial step will be the collection of different datasets from internet on various abusive comments perpetrated over social media internet which will be helpful in analyzing the abusive comments then those dataset will be aligned accordingly. Thenit will undergo a process called separation of datasets into training as well as testing where thetraining datasets will be used to train the model as well as testing will be used for evaluating the model. Then dataset pre-processing will be done which will align all the datasets into a specific category.

There exist several regression algorithms in machine learning to develop

an abusive comment detection model such as random forest regressor algorithm . In the proposed system, Random forest regressor, a regression machine learning algorithm is used to train the dataset consisting of data on the abusive comment detection and identification.By accurately detecting and marking abusive comments over social media it can be very helpful in improving the social media experience.

4.3 MODULE DESCRIPTION:

1. Dataset Collection

   ModuleSeparation

   of Dataset

   Labelling the Dataset Prior Training
2. Splitting of

   Dataset

   80%

   Training

   20%

   Testing
3. Dataset Pre-Processing Module
4. Training with Random Forest

   AlgorithmValidation and

   Evaluation

   Merging of data
5. Comment

   Detection

   Abusive

   Comment

   Non-Abusive Comment

4.3.1 DATASET COLLECTION MODULE:

A data set is a collection of data. Machine learning has become the go-to method for solving many challenging real-world problems. It’s definitely by far the best performing method for prediction tasks. These machine learning machines that have been working so well need fuel lots of fuel; that fuel is data. The more labelled data available, the better our model performs. The idea of more data leading to better performance has even been explored at a large-scale by Google with a dataset of 300 million images! When deploying a machine learning model in a real-world application**,** data must be constantly fed to continue improving

its performance. And, in the machine learning era, data is very well arguably the most valuable resource. There are three steps of collecting data.

Classification. When an algorithm to answer binary yes-or-no questions or to make a multi-class classification (grass, trees, or bushes; cats, dogs, or birds etc.)

Regression. For an algorithm to yield some numeric value. For example, if you spend too much time coming up with the right price for your product since it depends on many factors, regression algorithms can aid in estimating this value.

Ranking. Some machine learning algorithms just rank objects by a number of features. Ranking is actively used to recommend movies in video streaming services or show the products that a customer might purchase with a high probability based on his or her previous search and purchase activities.

Dataset Collection
Datas collected from Kaggle Website
These datas are used for Training using Machine Learning Algorithm

Figure 4.2 Dataset collection

4.3.2 SPLITTING OF DATASET

In machine learning, any dataset is usually split into two: training data and test data. The output variable along with other variables are included in the training set. The model learns the data and tries to generate some pattern. The other part of the dataset serves as atest set to validate our model’s prediction. The scikit library has a function called train_test_split to divide our data. Test size is the parameter which gives us the percentageof data that should belong to the test set. Train size stores the remaining part as the training dataset, either of which should be specified. Random state acts as a random number generator. For our dataset, we split the training and testing set with 80, 20 ratio the randomstate is passed as 0.

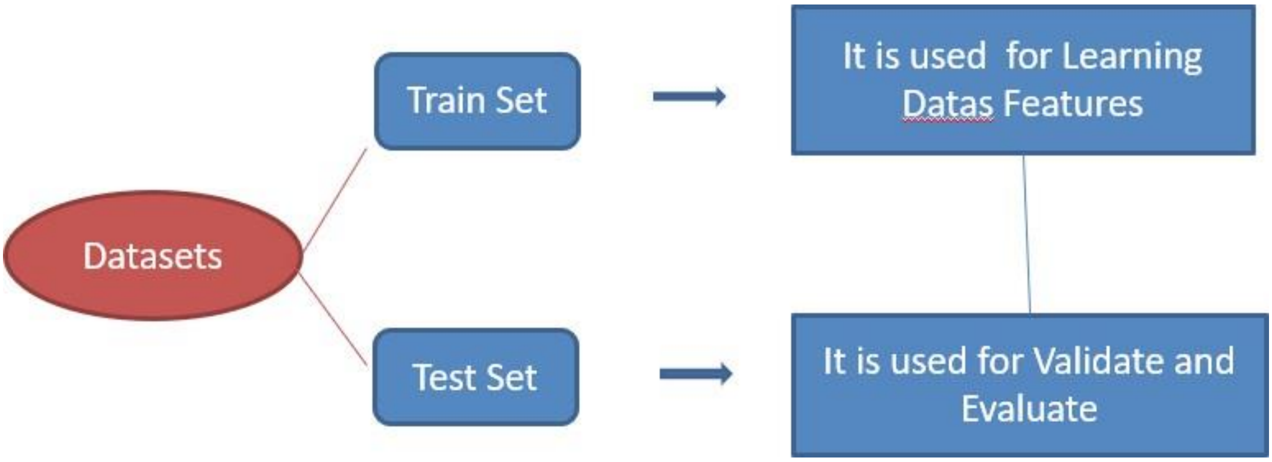

Train Set
It is used for Learning Datas Features
Datasets
Test Set
It is used for Validate and Evaluate

Figure 4.3 Splitting of Dataset

4.3.3 DATASET PRE-PROCESSING MODULE

Data pre-processing is a cleaning technique which is used to convert / transform theraw data into a clean and properly structured dataset suitable for further analysis. Data is usually collected and gathered from various sources, so it should be good enough and insome specific format before the model learns or gets trained with the data. This will help in achieving better and accurate results with valuable information. The basic steps in pre- processing involve filling up missing values and null values , getting rid of possible outliers and normalization.

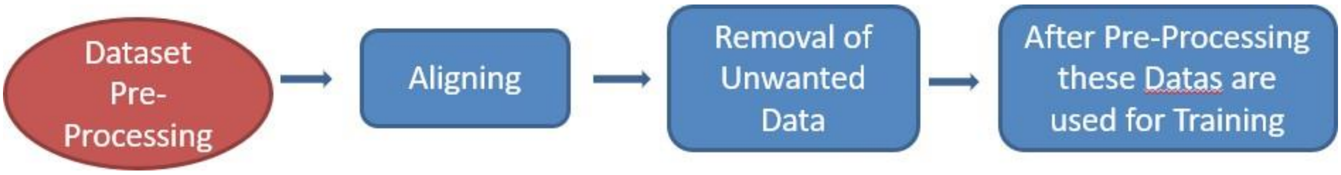


Figure 4.4 Dataset preprocessing

4.3.4 TRAINING WITH ALGORITHM

The Random Forest regressor is the regression machine learning algorithm to train the dataset consisting of abusive comments in social media. machine learning models require a lot of data in order for them to perform well. Usually, when training a machine learning model, one needs to collect a large, representative sample of data from a training set. Data from the training set can be as varied as a corpus of text, a collection of images, and data collected from individual users of a service. Overfitting is something to watch out for when training a machine learning model. Trained models derived from biased data can result in skewed or undesired predictions. Algorithmic bias is a potential result from data not fully prepared for training.

Random forest Algorithm

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks).

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. It technically is an ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly split dataset. This collection of decision tree classifiers is also known as the forest.

One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Random forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Fortunately, there's no need to combine a decision tree with a bagging classifier because you can easily use the classifier-class of random forest.

Below you can see how a random forest would look like with two trees:

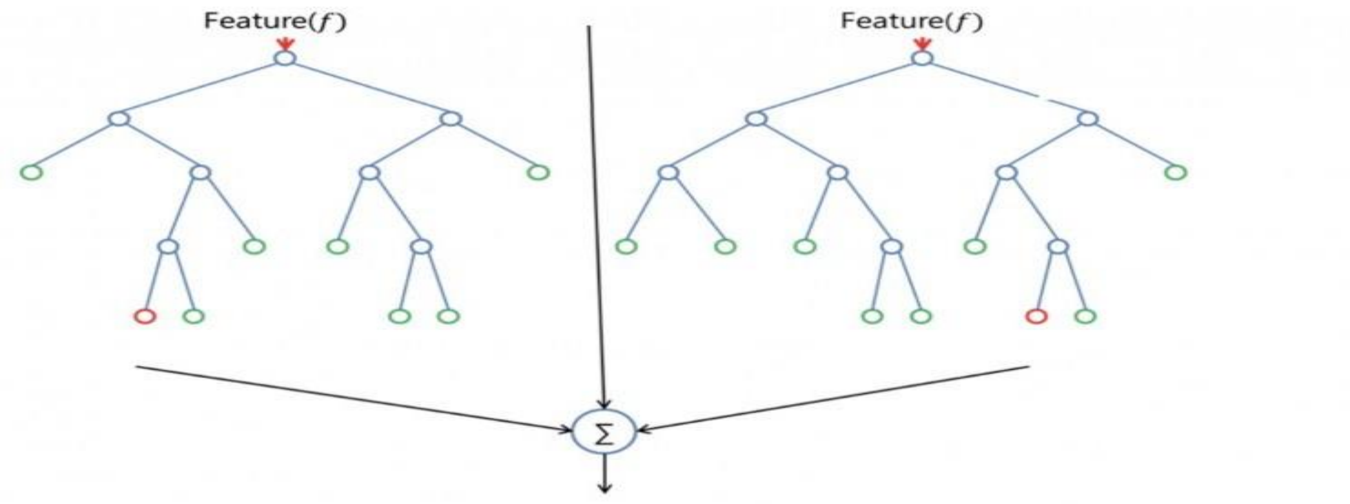


Figure 4.5 Random Forest regressor

Random Forest Algorithm Works in Four steps:

1.Select Random Samples from a given Datasets.

2.It will construct a Decision Tree for each sample and get a prediction result from eachDecision Tree.

3.It will Perform a majority vote for

predicted result.4.It will select the most votes

as the final prediction.

CHAPTER 5

SOFTWARE DESCRIPTION

5.1 Jupyter notebook

In this project the jupyter notebook is used as an IDE.

In this case, "notebook" or "notebook documents" denote documents that contain bothcode and rich text elements, such as figures, links, equations, ... Because of the mix of code and text elements, these documents are the ideal place to bring together an analysis description, and its results, as well as, they can be executed perform the data analysis in real time.

At some point, we all need to show our work. Most programming work is shared either as raw source code or as a co mpiled executable. The source code provides complete information, but in a way that’s more “tell” than “show.” The executable shows us what the software does, but even when shipped with the source code it can be difficult to grasp exactlyhow it works.

A notebook integrates code and its output into a single document that combines visualizations, narrative text, mathematical equations, and other rich media. In other words: it's a single document where you can run code, display the output, and also add explanations, formulas, charts, and make your work more transparent, understandable, repeatable, and shareable.

Using Notebooks is now a major part of the data science workflow at companies acrossthe globe. If your goal is to work with data, using a Notebook will speed up your workflow and make it easier to communicate and share your results.

Imagine being able to view the code and execute it in the same UI, so that you could make changes to the code and view the results of those changes instantly, in real time? That’sjust what Jupyter Notebook offers.

Jupyter Notebook was created to make it easier to show one’s programming work, and to let others join in. Jupyter Notebook allows you to combine code, comments, multimedia, and visualizations in an interactive document — called a notebook, naturally — that can be shared, re-used, and re-worked.

And because Jupyter Notebook runs via a web browser, the notebook itself could be hosted on your local machine or on a remote server

One major feature of the Jupyter notebook is the ability to display plots that are the output of running code cells. The IPython kernel is designed to work seamlessly with the matplotlib plotting library to provide this functionality. Specific plotting library integration is a feature of the kernel..

Each **.ipynb** file is one notebook, so each time you create a new notebook, anew **.ipynb** file will be created.

Each **.ipynb** file is a text file that describes the contents of your notebook in a format called JSON. Each cell and its contents, including image attachments that have been converted into strings of text, is listed therein along with some metadata.

Jupyter Notebooks are a powerful way to write and iterate on your Python code for data analysis. Rather than writing and re-writing an entire program, you can write lines of codeand run them one at a time. Then, if you need to make a change, you can go back and make your edit and rerun the program again, all in the same window.

Jupyter Notebook is built off of IPython, an interactive way of running Python code in the terminal using the REPL model (Read-Eval-Print-Loop). The IPython Kernel runs the computations and communicates with the Jupyter Notebook front-end interface. It also allows Jupyter Notebook to support multiple languages. Jupyter Notebooks extend IPython through additional features, like storing your code and output and allowing you to keep markdown notes.

Jupyter Notebook provides you with an easy-to-use, interactive data science environment across many programming languages that doesn't only work as an IDE, but also as a presentation or education tool.

5.2 Python

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.

Python is Interactive − You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python is Object-Oriented − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.

Python is a Beginner's Language − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

In this project python is used as programming language.

In technical terms, Python is an object-oriented, high-level programming languagewith integrated dynamic semantics primarily for web and app development. It is extremely attractive in the field of Rapid Application Development because it offers dynamic typing and dynamic binding options.

Python is relatively simple, so it's easy to learn since it requires a unique syntax that focuses on readability. Developers can read and translate Python code much easier than other languages. In turn, this reduces the cost of program maintenance and development because it allows teams to work collaboratively without significant language and experience barriers.

Additionally, Python supports the use of modules and packages, which means that programs can be designed in a modular style and code can be reused across a variety of projects. Once you've developed a module or package you need, it can be scaled for use in other projects, and it's easy to import or export these modules.

One of the most promising benefits of Python is that both the standard library and the interpreter are available free of charge, in both binary and source form. There is no exclusivity either, as Python and all the necessary tools are available on all major platforms. Therefore, it is an enticing option for developers who don't want to worry about paying high development costs.

If this description of Python over your head, don't worry. You'll understand it soon enough. What you need to take away from this section is that Python is a programming language used to develop software on the web and in app form, including mobile. It'srelatively easy to learn, and the necessary tools are available to all free of charge.

import pandas as pd

import pandas as pd. Simply imports the library that current namespace, but rather thanusing the name pandas , it's instructed to use the name pd instead. This is just so you can do pd. whatever instead of having to type out pandas. whatever all the time if you just do import pandas.

import numpy as np

NumPy is an open-source numerical Python library. NumPy contains a multi-dimensional array and matrix data structures. It can be utilised to perform a number of mathematical operations on arrays such as trigonometric, statistical, and algebraic routines. NumPy **is** an extension of Numeric and Numarray.

import Random

import random imports the random module, which contains a variety of things to do with random number generation. Among these is the random() function, which generates random numbers between 0 and 1.

import matplotlib.pyplot as plt

Pyplot is a collection of functions in the popular visualization package Matplotlib. Its functions manipulate elements of a figure, such as creating a figure, creating a plotting area, plotting lines, adding plot labels, etc.

import seaborn as sns

Seaborn helps you explore and understand your data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

Sklearn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. It is a free software machine learning library for the Python programming language.

sklearn.metrics

Classification metrics. The sklearn. metrics module implements several loss, score, andutility functions to measure classification performance.Some metrics might requireprobability estimates of the positive class, confidence values, or binary decisions values. **import roc_curve**

ROC curves typically feature true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the plot is the “ideal” point - a false positive rateof zero, and a true positive rate of one. This is not very realistic, but it does mean that a largerarea under the curve (AUC) is usually better.

The steepness of ROC curves is also important, since it is ideal to maximize the true positive rate while minimizing the false positive rate.

ROC curves are typically used in binary classification to study the output of a classifier. In order to extend ROC curve and ROC area to multi-label classification, it is necessary to binarize the output. One ROC curve can be drawn per label, but one can also draw a ROC curve by considering each element of the label indicator matrix as a binary prediction.

CHAPTER 6
SOURCE CODE

Dataset

Collection code

```
import pandas as
pd import
numpy as np
import random
from sklearn.feature_extraction.text import
CountVectorizerfrom sklearn.feature_extraction.text
import TfidfVectorizer
from sklearn.model_selection import
train_test_split,GridSearchCVimport matplotlib.pyplot as
plt
import seaborn as sns
label_data =
pd.read_csv("bullying_dataset.csv")
type(label_data)
label_data.head()
```

Preprocessing Dataset

```
 y = label_data["label"]
label_data1 =
label_data["tweet"] def
makeTokens(f):
   tkns_BySlash = str(f.encode('utf-8')).split('/')    # make tokens after splitting
   by slashtotal_Tokens = []
   for i in tkns_BySlash:
      tokens = str(i).split('-')     # make tokens after splitting
      by dashtkns_ByDot = []
      for j in range(0,len(tokens)):
         temp_Tokens = str(tokens[j]).split('.')     # make tokens after
         splitting by dottkns_ByDot = tkns_ByDot + temp_Tokens
      total_Tokens = total_Tokens + tokens + tkns_ByDot
   total_Tokens = list(set(total_Tokens))    #remove
   redundant tokensreturn total_Tokens
 vectorizer =
 CountVectorizer(tokenizer=makeTokens)X =
 vectorizer.fit_transform(label_data1)
from collections import Counter
    from sklearn.preprocessing import
    LabelEncoderfrom keras.utils import
    np_utils
    def clean_str(string):
    #print (string)
    return string.strip().lower()
```

Random Forest Algorithm Traning and Testing Datasets

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)print(X_train[0])
from sklearn.ensemble import
RandomForestClassifierrtree =
RandomForestClassifier()
rtree.fit(X,y)
from sklearn.metrics import
classification_reportY_rt_predict =
rtree.predict(X_test)
print(classification_report(y_test,Y_rt_pr
edict))
```

Confusion Matrix Using Random Forest Classifier

```
from sklearn.metrics import confusion_matrix
confusion_matrix1 = pd.DataFrame(confusion_matrix(y_test,
Y_rt_predict))plt.figure()
plt.title('Confusion Matrix using RandomForest Classifier')
sns.heatmap(confusion_matrix1,annot=True,cmap='Greens',fmt='.2f')
```

Graph Obtained Using Random Forest Classifier

```
from sklearn.metrics import
roc_auc_scorefrom sklearn.metrics
import roc_curve
logit_roc_auc2 = roc_auc_score(y_test,
Y_rt_predict)plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False
Positive Rate')
plt.ylabel('True
Positive Rate')
plt.title('Receiver operating
characteristic')#Random Forest
Curve
fpr, tpr, thresholds = roc_curve(y_test,
rtree.predict_proba(X_test)[:,1]) plt.plot(fpr, tpr, label='Random
Forest (area = %0.2f)' % logit_roc_auc2)plt.legend(loc="lower
right")
plt.show()
```

CHAPTER – 7
TESTING

7.1 TEST PROCEDURE

Testing is performed to identify errors. It is used for quality assurance. Testing is an integralpart of the entire development and maintenance process. The goal of the testing during phaseis to verify that the specification has been accurately and completely incorporated into thedesign, as well as to ensure the correctness of the design itself. For example, the design mustnot have any logic faults in it. If it is not detected before coding commences, the cost offixing the faults will be considerably higher as reflected. Detection of design faults can be achieved by means of inspection as well as walkthrough. Testing is one of the important stepsin the software development phase.

7.2 MANUAL TESTING

Manual Testing is a type of software testing in which test cases are executed manually by a tester without using any automated tools. The purpose of Manual Testing is to identify the bugs, issues, and defects in the software application. Manual software testing is the most primitive technique of all testing types and it helps to find critical bugs in the software application.

Any new application must be manually tested before its testing can be automated. Manual Software Testing requires more effort but is necessary to check automation feasibility. Manual Testing concepts does not require knowledge of any testing tool.

CHAPTER 8

RESULTS AND DISCUSSION

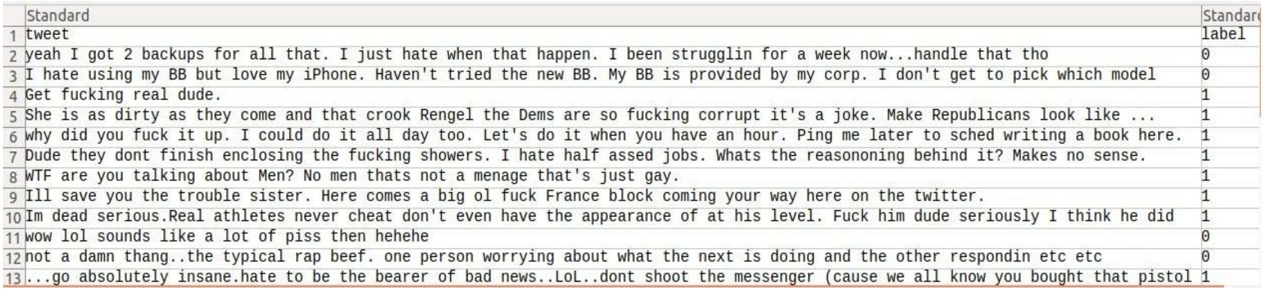
8.1 FINAL RESULTS OBTAINED:

To begin with, testing of the trained model, we can split our project into modules ofimplementation that is done.

Dataset collection involves the process of collecting abusive comments from variousinternet sources and social media.

Various datasets were collected and one example among the collected dataset can befound below:

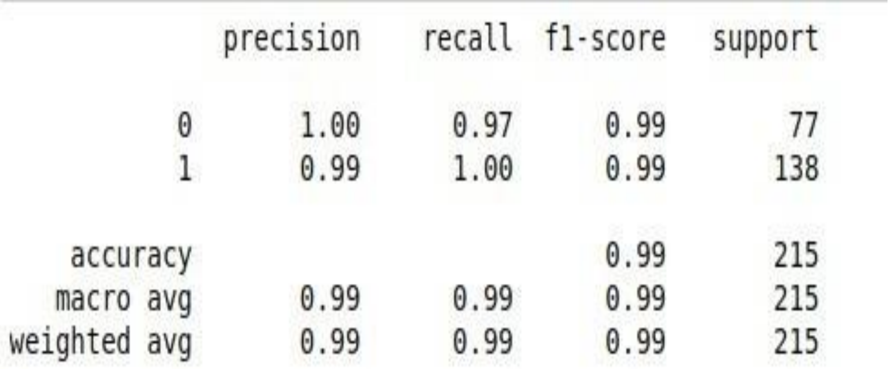
The below screenshot shows a sample of dataset collected:



	Standard	Standar
1	tweet	label
2	yeah I got 2 backups for all that. I just hate when that happen. I been strugglin for a week now...handle that tho	0
3	I hate using my BB but love my iPhone. Haven't tried the new BB. My BB is provided by my corp. I don't get to pick which model	0
4	Get fucking real dude.	1
5	She is as dirty as they come and that crook Rengel the Dems are so fucking corrupt it's a joke. Make Republicans look like ...	1
6	why did you fuck it up. I could do it all day too. Let's do it when you have an hour. Ping me later to sched writing a book here.	1
7	Dude they dont finish enclosing the fucking showers. I hate half assed jobs. Whats the reasononing behind it? Makes no sense.	1
8	WTF are you talking about Men? No men thats not a menage that's just gay.	1
9	Ill save you the trouble sister. Here comes a big ol fuck France block coming your way here on the twitter.	1
10	Im dead serious.Real athletes never cheat don't even have the appearance of at his level. Fuck him dude seriously I think he did	1
11	wow lol sounds like a lot of piss then hehehe	0
12	not a damn thang..the typical rap beef. one person worrying about what the next is doing and the other respondin etc etc	0
13	...go absolutely insane.hate to be the bearer of bad news..LoL..dont shoot the messenger (cause we all know you bought that pistol	1

Figure 8.1 Dataset collection

The below image shows the training using random forest classifier:



```
              precision    recall  f1-score   support

           0       1.00      0.97      0.99        77
           1       0.99      1.00      0.99       138

    accuracy                           0.99       215
   macro avg       0.99      0.99      0.99       215
weighted avg       0.99      0.99      0.99       215
```

Figure 8.2 Training using random forest classifier

A confusion matrix is a table that is often used to describe the performance of aclassification model (or "classifier") on a set of test data for which the true values are known. The below image shows the confusion matrix using random forest classifier:

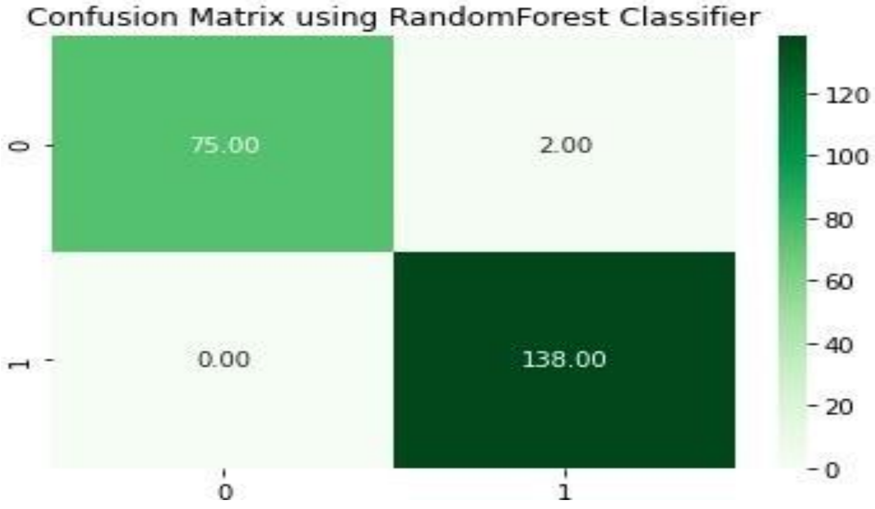


Figure 8.3 Confusion matrix using random forest

classifier the below image shows the graph obtained using random

forest classifier:

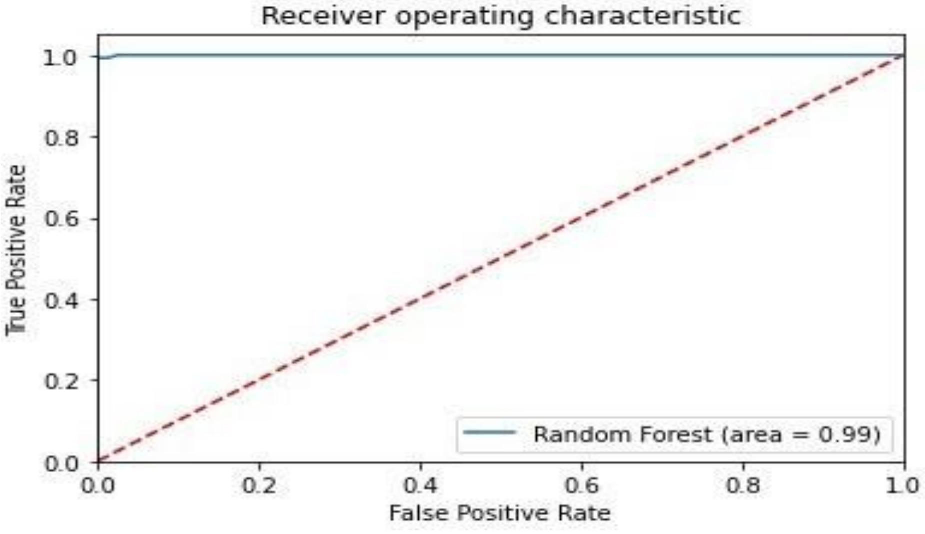
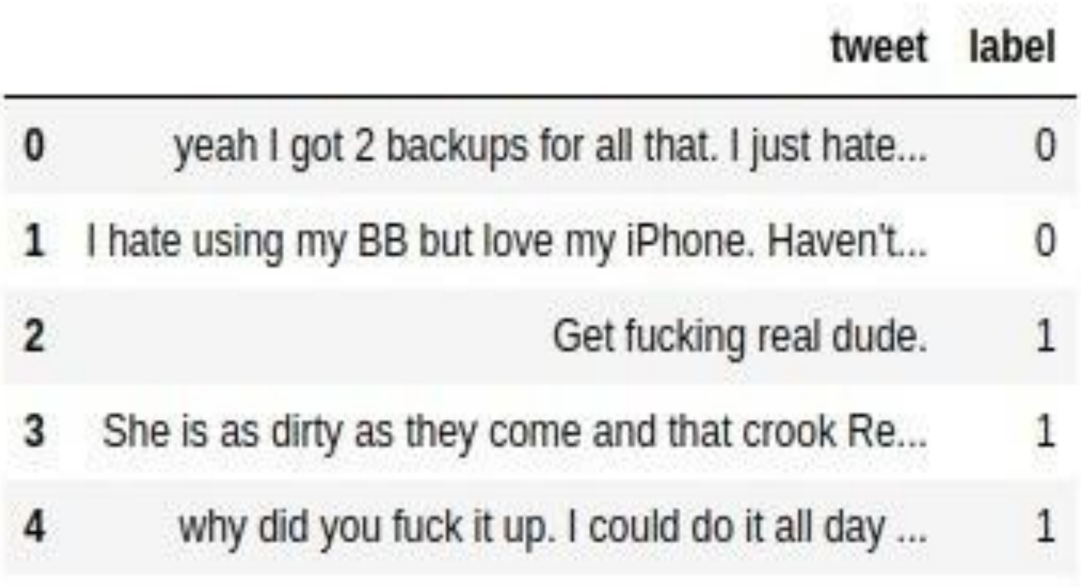


Figure 8.4 Graph obtained using random forest classifier

The below image shows the effective detection and prediction of abusive comments:



	tweet	label
0	yeah I got 2 backups for all that. I just hate...	0
1	I hate using my BB but love my iPhone. Haven't...	0
2	Get fucking real dude.	1
3	She is as dirty as they come and that crook Re...	1
4	why did you fuck it up. I could do it all day ...	1

Figure 8.5 Abusive comment detection and prediction

CHAPTER - 9

CONCLUSION AND FUTURE WORK

9.1 CONCLUSION

In this project, we have successfully implemented a system of effective detection and identification of abusive comments over social media. The abusive comments dataset is trained using random forest classifier, a machine-learning algorithm. The present detection methods are inaccurate and inefficient. The system has provided an easy and efficient solutionat very cost-effective approach.

Hence, The goal of our project is to filter tweets or comments using machine learning algorithms and to create an environment which is healthy for interaction and transfer of information between individuals.

We as the future generation are responsible for preventing the ongoing toxic environment in social media and create a healthy environment in social media.

9.2 FUTURE WORK

In the coming future, we review this application of abusive comment identification model to identification and detection of the abusive comment with more accuracy and efficiency.

The application has good use in the social media. In the social media space, they have more chance to develop or convert this project in many ways. Thus, this project has an efficient scope in coming future where manual detection and prediction can be converted to computerized prediction in a cheap way.

CHAPTER 10

REFERENCES

[1] Yuan Gao, Maoguo Gong, Senior Member, IEEE, Yu Xie, and A. K. Qin, Senior Member, IEEE,"An attention-based unsupervised adversarial model for review spam detection"[2020]

[2] An Improved Rainfall Forecasting Model Based on GNSS Observations, Qingzhi Zhao , Yang Liu , Xiongwei Ma , Wanqiang Yao, Yibin Yao , and Xin Li,[2020]

[3] An Ensemble Forecasting Method for the Aggregated Load With Subprofiles, Yi Wang; Qixin Chen; Mingyang Sun; Chongqing Kang; Qing Xia,[2019]

[4] Probabilistic Load Forecasting via Quantile Regression Averaging on Sister Forecasts, Bidong Liu; Jakub Nowotarski; Tao Hong; Rafał Weron,[2018]

[5] Long-Term Retail Energy Forecasting With Consideration of Residential Customer Attrition,Jingrui Xie SAS Institute,Cary,NC,USA; Tao Hong;Joshua Stroud,[2017]

[6] Bala Sundara Ganapathy.N,Helda Mercy.M,Giftson Vasanth Samuel Raj. A,"A Framework for Social Media Network to Curtail the Banned and Abused Images ",[2017]

[7] Han Hu, Pranavi Moturu, Kannan Neten Dharan, James Geller, Sophie Di Iorio, Hai Phan,"Deep Learning Model for Classifying Drug Abuse Risk Behavior in Tweets"[2016]

[8] Sayeed Ahsan Khan,Mohammed Hazim Alkawaz,Hewa Majeed Zangana,"The Use and Abuse of Social Media for Spreading Fake News"[2016]

[9] LAIHANG YU1 , (Student Member, IEEE), LIN FENG2 , CHEN CHEN3 , (Member, IEEE), TIE QIU4 , (Senior Member, IEEE), LI LI1 , AND JUN WU2,"A Novel Multi- Feature Representation of Images for Heterogeneous IoTs"[2015]

[10] Ryan Sequeira , Avijit Gayen , Niloy Ganguly, Senior Member, IEEE, Sourav Kumar Dandapat, and Joydeep Chandra,"A Large-Scale Study of the Twitter Follower Network to Characterize the Spread of Prescription Drug Abuse Tweets"[2018]

[11] Semiu Salawu, Yulan He, and Joanna Lumsden,”Approaches to Automated Detection ofCyberbullying: A Survey ”[2014]

[12] Yubao Zhang, Student Member, IEEE, Xin Ruan, Student Member, IEEE, Haining Wang, Senior Member, IEEE, Hui Wang, and Su He,”Twitter Trends Manipulation: A First Look Inside the Security of Twitter Trending”[2013]

[13] Ryan Sequeira , Avijit Gayen , Niloy Ganguly, Senior Member, IEEE, Sourav Kumar Dandapat, and Joydeep Chandra,”A Large-Scale Study of the Twitter Follower Network to Characterize the Spread of Prescription Drug Abuse Tweets”[2012]

[14] Gaoyang Liu, Chen Wang, Kai Peng, Haojun Huang, Yutong Li, Wenqing Cheng,” SocInf: Membership Inference Attacks on Social Media Health Data With Machine Learning”[2011]

[15] J. Jiménez, K. Donado and C. G. Quintero,” A Methodology for Short-Term Load Forecasting”

[16] Mao Tan, Member, IEEE, Siping Yuan, Shuaihu Li, Yongxin Su, Hui Li, and Feng He,”Ultra-short-term industrial power demand forecasting using LSTM based hybrid ensemble learning”[2010]

[17] Chinnawat Surussavadee, Senior Member, IEEE,” Evaluation of High-Resolution Tropical Weather Forecasts Using Satellite Passive Millimeter-Wave Observations”[2009]

[18] Ming Yang, Member, IEEE, You Lin, Student Member, IEEE, and Xueshan Han,” Probabilistic Wind Generation Forecast Based on Sparse Bayesian Classification and Dempster-Shafer Theory”