CYBERBULLYING DETECTION IN TWITTER CLONE USING MACHINE LEARNING ALGORITHM

ST. TERESA'S COLLEGE(AUTONOMOUS)

AFFILIATED TO MAHATMA GANDHI UNIVERSITY

PROJECT REPORT

In partial fulfilment of the requirements for the award of the degree of

BCA (CLOUD TECHNOLOGY & INFORMATION SECURITY MANAGEMENT)

By

DEVIKA PRASANNAKUMAR- SB21BCA010

&

SHEHERNA SHAJAHAN- SB21BCA032

III DC BCA (CLOUD TECHNOLOGY & INFORMATION SECURITY MANAGEMENT)

Under the guidance of

Mrs. KAVYA KISHORE

DEPARTMENT OF BCA(CLOUD TECHNOLOGY & INFORMATION SECURITY MANAGEMENT)

MARCH 2024

CYBERBULLYING DETECTION IN TWITTER CLONE USING MACHINE LEARNING ALGORITHM

ST. TERESA'S COLLEGE(AUTONOMOUS)

AFFILIATED TO MAHATMA GANDHI UNIVERSITY

PROJECT REPORT

In partial fulfillment of the requirements for the award of the degree of

BCA (CLOUD TECHNOLOGY & INFORMATION SECURITY MANAGEMENT)

By

DEVIKA PRASANNAKUMAR- SB21BCA010

&

SHEHERNA SHAJAHAN- SB21BCA032

III DC BCA (CLOUD TECHNOLOGY & INFORMATION SECURITY MANAGEMENT)

Under the guidance of

Mrs. KAVYA KISHORE

DEPARTMENT OF BCA(CLOUD TECHNOLOGY & INFORMATION SECURITY MANAGEMENT)

MARCH 2024

DECLARATION

We, undersigned, hereby declare that the project report, **'CYBERBULLYING DETECTION IN TWITTER CLONE USING MACHINE LEARNING ALGORITHM**" ,submitted for partial fulfillment of the requirements for the award of degree of BCA (Cloud Technology & Information Security Management)at St. Teresa's College (Autonomous),Ernakulam (Affiliated to Mahatma Gandhi University), Kerala, is a bonafide work done by us under the supervision of Mrs. Kavya Kishore. This submission represents our ideas in our own words and where ideas or words of others have not been included. We have adequately and accurately cited and referenced the original sources. We also declare that we have adhered to the ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in our submission. We understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Ernakulam
March 2024

DEVIKA PRASANNAKUMAR – SB21BCA010
SHEHERNA SHAJAHAN – SB21BCA032

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

BCA (CLOUD TECHNOLOGY & INFORMATION

SECURITY MANAGEMENT)

DEPARTMENT OF BCA(CLOUD TECHNOLOGY & INFORMATION SECURITY MANAGEMENT)

CERTIFICATE

This is to certify that the report entitled "**CYBERBULLYING DETECTION IN TWITTER CLONE USING MACHINE LEARNING ALGORITHM**", submitted by DEVIKA PRASANNAKUMAR AND SHEHERNA SHAJAHAN to the Mahatma Gandhi University in partial fulfillment of the requirements for the award of the Degree of BCA (Cloud Technology & Information Security Management) is a bonafide record of the project work carried out by them under our guidance and supervision. This report in any formhas not been submitted to any other University or Institute for any purpose.

ARCHANA P MENON

Head of the department

Mrs. KAVYA KISHORE

Internal Supervisor **External Supervisor**

ACKNOWLEDGEMENT

First and foremost we thank God Almighty for his blessings. We take this opportunity to express our gratitude to all those who helped us in completing this project successfully. I wishto express our sincere gratitude to the **Manager Rev. Dr. Sr. Vinitha CSST** and the Principal**Dr. Lizzy Mathew** for providing all the facilities.

We express our sincere gratitude towards the Head of the department **Mrs. Archana P Menon** for the support. We deeply express sincere thanks to our project guide **Mrs. Kavya Kishore** for her proper guidance and support throughout the project work.

We are indebted to our beloved teachers whose cooperation and suggestion throughout the project helped us a lot. We thank all our friends and classmates for their support.

We convey our hearty thanks to our parents for the moral support, suggestion and encouragement.

ABSTRACT

Cyber bullying is a serious social issue and it has been increased since the inceptionof social media such as Twitter and Facebook. Cyberbullying can have serious legal consequences in many world countries. So, Detecting and blocking such posts from the feed are really important and many social media platforms are developing their own technologies to defend against cyber bullying. There have been many approaches from simple text-based methods to advanced AI/ML techniques to fight cyberbullyingtill date. This project is to develop a real-time cyber bullying detection clone using HTML and CSS with a well-trained Machine Learning model at the back-end to detect cyber bullying from text posts. For training the MLmodel, text data of both cyber bullying and normal twitter posts are collected from various sources on the internet. This data is passed through a series of processes called text data pre- processing. We will use this dataset to train a number of Machine Learning algorithms namely Decision Tree, Random forest and XGBoost, etc. After training, these models are tested to detect cyber bullying in text posts and the best performing model will be finalized. The back-end part which is the data collection, data analysis, data preparation, and machine learning model building and training are done in Google Colab platform in Python 3.7. The front-end which is a web app and a clone of the twitter itself displaying the result is developed in HTML and CSS and locally hosted with the help of django framework.

TABLE OF CONTENTS

LIST OF FIGURES ..... **i**

LIST OF ABBREVIATIONS ..... **ii**

Chapter 1 INTRODUCTION **1**

HISTORY OF CYBERBULLYING 10

CYBERBULLYING IN INDIA 19

TWITTER CYBERBULLYING 21

HOW TWITTER IS FIGHTING HARASSMENT & CYBERBULLYING 23

Chapter 2 LITERATURE SURVEY… **25**

Chapter 3 EXISTING SYSTEM **32**

RULE-BASED SYSTEMS 32

KEYWORD FILTERING 33

SENTIMENT ANALYSIS 34

SOCIAL NETWORK ANALYSIS 34

HUMAN MODERATION 35

HYBRID APPROACHES 36

Chapter 4 PROPOSED SYSTEM **38**

Chapter 5 SYSTEM DESIGN AND ARCHITECTURE… **40**

THE MACHINE LEARNING MODELS 45

DECISION TREE CLASSIFIER 45

RANDOM FOREST CLASSIFIER 46

XGBOOST ALGORITHM 48

Chapter 6 SYSTEM REQUIREMENTS **50**

Chapter 7 MODULE DESCRIPTION… **50**

THE DATA COLLECTION AND PREPROCESSING MODULE 51

THE MACHINE LEARNING MODULE 53

THE APPLICATION MODULE 54

Chapter 8 IMPLEMENTATION… **55**

DATA COLLECTION AND PREPROCESSING 55
THE MACHINE LEARNING MODELS 58

Chapter 9 CONCLUSION 62

REFERENCES

APPENDICES

LIST OF FIGURES

4.1 The proposed system DFD 40
4.2 The use case diagram 41
5.1 System Architecture 42
5.2 Decision Tree 46
5.3Random forest 47
8.1 Types of tweets in the dataset 57
8.2 Word distribution in age category 58
8.3 Word distribution in ethnicity category 58
8.4 Word distribution in gender category 58
8.5 Word distribution in age category 58
8.6 Word distribution in non bullying category 58
8.7 Word distribution in other bullying categories 58
8.8 Test result 1 60
8.9 Test result 2 60
8.10 The landing page 61
8.11 The login page 61
8.12 The main page 62
8.13 The Final Test Result 63
8.14 The Final Test Result 2 64

Chapter 1
INTRODUCTION

Cyberbullying is bullying with the use of digital technologies. It can take place on social media, messaging platforms, gaming platforms and mobile phones. It is repeated behavior, aimed at scaring, angering or shaming those who are targeted. Examples include:

- spreading lies about or posting embarrassing photos or videos of someone on social media
- sending hurtful, abusive or threatening messages, images or videos via messaging platforms
- impersonating someone and sending mean messages to others on their behalf or through fake accounts.

Face-to-face bullying and cyberbullying can often happen alongside each other. But cyberbullying leaves a digital footprint – a record that can prove useful and provide evidence to help stop the abuse. Cyberbullying also includes photos, messages, or pages that don't get taken down, even after the person has been asked to do so. In other words, it's anything that gets posted online and is meant to hurt, harass, or upset someone else. Intimidation or mean comments that focus on things like a person's gender, religion, sexual orientation, race, or physical differences count as discrimination, which is against the law in many states. That means the police could get involved, and bullies may face serious penalties. Online bullying can be particularly damaging and upsetting because it's usually anonymous or hard to trace. It's also hard to control, and the person being victimized has no idea how many people (or hundreds of people) have seen the messages or posts. People can be tormented non stop whenever they check their device or computer. Online bullying and harassment can be easier to commit than other acts of bullying because the bully doesn't have to confront their target inperson.

Cyberbullying is one of the many cons that come with internet usage. Behaviors that are considered to be cyberbullying include, but are not limited, to:

- Exclusion
- Harassment
- Outing
- Derogatory language
- Cyberstalking Impersonation

- Dissing (insulting)
- 'Trolling' (intentionally provoking a negative response)
- 'Catfishing' (using fake profiles to deceive others)

How to respond to cyberbullying?

Sometimes, people are afraid or not sure if they're being bullied or not. So they don't do anything about it. If you're being bullied, harassed, or teased in a hurtful way or know someone who is, you don't have to suffer in silence. In fact, you absolutely should report any upsetting texts, messages, posts, or emails.

Tell someone.

Most experts agree: The first thing to do is tell an adult you trust. This is often easier said than done. People who are cyberbullied may feel embarrassed or reluctant to report a bully. Some may hesitate because they're not 100% sure who is doing the bullying. But bullying can get worse, so speak up until you find someone to help. Sometimes the police can track down an anonymous online bully, so it's often worthwhile to report it. Most parents are so concerned about protecting their kids that sometimes they focus on taking major steps to stop the bullying. If you're being bullied and worry about losing your phone or computer privileges, explain your fears to your parents. Let them know how important it is to stay connected, and work with them to figure out a solution that doesn't leave you feeling punished as well as picked on. You may have to do some negotiating on safe phone or computer use — the most important thing is to first get the bullying under control.

Walk away.

What you've heard about walking away from a real-life bully works in the virtual world too. Ignoring bullies is the best way to take away their power, but it isn't always easy to do — in the real world or online. If you see something upsetting, try to step away from the computer or turn off your phone for a while. Don't respond, and never forward the message to someone else. Find something to distract yourself from what's going on. Do something you love that doesn't give you time to think about what's happening, like playing the guitar, going for a run, or losing yourself in a book or movie. You can also just chat with a parent or sibling or play with a pet. Taking a break like this allows you to keep things in perspective and focus on the good things in your life. It also gives you time to figure out how you want to handle things.

Resist the urge to retaliate or respond.

Walking away or taking a break when you're faced with online bullying gives you some spaceso you won't be tempted to fire back a response or engage with the bully or bullies. Responding when we're upset can make things worse. (Standing up to a bully can be effective sometimes, but it's more likely to provoke the person and escalate the situation.) Taking a break gives the power back to you! Although it's not a good idea to respond to a bully, it is a good idea to save evidence of the bullying if you can. It can help you prove your case, if needed. You don't have to keep mean emails, texts, or other communications where you see them all the time. You can screenshot them or ask a parent to make a copy or save them to a flash drive.

Report bullying.

Social media sites take it seriously when people post cruel or mean stuff or set up fake accounts. If users report abuse, the site administrator may block the bully from using the site in the future. If someone sends you mean texts or emails, report it to phone service or email providers (such as Comcast, Google, and Verizon).

Block the bully.

Most devices have settings that let you electronically block the bully or bullies from sending notes. If you don't know how to do this, ask a friend or adult who does.

Be safe online.

Passwords protect your smartphone and your online sites, and change your passwords often. Be sure to share your passwords only with your parents or guardian. It's also wise to think twice before sharing personal information or photos/videos that you don't want the world to see. Once you've posted a photo or message, it can be hard or even impossible to delete. So remind yourself to be cautious when posting photos or responding to someone's upsetting message.

History of Cyberbullying:

Traditional bullying forced its way onto the web in the 1990s with the advent of affordable personal computers. Since then, classmates (and even strangers) have subjected children and

teens to cyberbullying in public chat rooms or on private messaging platforms. The web's anonymity provided the perfect cover for a user to harass or intimidate others without many repercussions. It's easy to assume that online teasing isn't as harmful as in-person bullying, but that certainly isn't the case. And although several US states have enacted laws in recent years to regulate teen cyberbullying, the wider-reaching effects can be harmful or even deadly. In this post, we trace the history of cyberbullying and how the issue has evolved so that you can be better prepared to help protect your kids from being bullied online.

Cyberbullying On the Books

In response to the 1999 Columbine school shooting, states began to pass anti-bullying laws. Some of these laws included cyberbullying as an offense, but many did not. Cyberbullying was brought to the mainstream after online harassment resulted in multiple teen suicides. Oneof the earliest cases occurred in 2007 when 13-year-old Megan Meier died by suicide after neighbors created a fake Myspace profile under the name "Josh Evans" to harass her. A federal grand jury found the perpetrators guilty of conspiracy and unauthorized computer use, but they were later acquitted. Meier's case spurred her home state of Missouri to pass an anti- harassment law encompassing acts of cyberbullying.

The Internet Gets Mobile

Cyberbullying hit its stride in the mid-2000s when smartphones became the newest must- have piece of technology. Teens could now share text messages and photos with unprecedented ease. An 18-year-old named Jessica Logan killed herself after her boyfriend sent nude photos of her to teenagers in at least seven Ohio high schools. Logan was then relentlessly cyberbullied through Myspace and text messages. A year later, in a nearly identical case, a 13-year-old named Hope Sitwell killed herself after her boyfriend sent a nude photo of her to students in six high schools in Florida. Both deaths resulted in lawsuits against the schools as well as new state cyberbullying laws.

While social networks have undoubtedly led to the largest increase in cyberbullying, the problem dates back to the 1990s and some of the Internet's first forums. It followed the mobile revolution as text messages and mobile internet became commonplace.

Statistics of cyberbullying.

Internet connectivity is important because it provides both educational and social benefits for young people. Unfortunately, these positive attributes are counterbalanced by potentially

dangerous consequences. Alongside improving communication and democratizing access to information, the internet lets people conceal themselves behind a mask of anonymity. This creates a whole new set of risks for children – and often adults too. The internet creates not only a threat for teens who could fall victim to cyberbullying – but also the potential for children to engage in online crimes, trolling, and cyberbullying themselves. That makes cyberbullying a topic that all parents and guardians need to be aware of.

In total 20,793 interviews were conducted between March 23 – April 6, 2018, among adults aged 18-64 in the US and Canada, and adults aged 16-64 in all other countries. Of particular interest are Russia and Japan. In both countries, parents expressed extremely high levels of confidence that their children did not experience cyberbullying of any kind. Meanwhile, Indian parents remained among the highest to express confidence that their children were cyberbullied at least sometimes, a number that only grew from 2011 to 2018. Across Europe and the Americas, it also appears more parents are either becoming aware of their children's negative experiences with cyberbullying, or their children are increasingly experiencing such attacks online.

One-fifth of all bullying occurs through social media.

Although the vast majority of parents reported bullying in school, 19.2% stated that bullying occurred through social media sites and apps. A further 11% indicated bullying occurred through text messages, while 7.9% identified video games as a source. Meanwhile, 6.8% reported bullying occurred on non-social media websites, while 3.3% indicated the bullying occurred through email. Some parents even witnessed cyberbullying occur, with 10.5% of parents indicating they observed the cyberbullying themselves. Interestingly, subsequent research has indicated that you don't even have to have internet access to be affected - Guimetti et al. (2020) found a positive correlation between time spent using a cellphone (but not online) and the likelihood of cyberbullying victimization. A study written by scholars working at the Universities of Florida and Denver revealed that the global pandemic had a marked effect on cyberbullying levels on Twitter. According to that study, the analysis of 454,046 publicly available tweets related to cyberbullying revealed a direct correlation between the pandemic and cyberbullying incidents. According to another study by L1GHT, a company that specializes in AI that is designed to detect and filter toxic content to protect children, online toxicity and cyberbullying on social media sites and video conferencing apps increased by up to 70% (PDF) due to the pandemic. This included a 200% increase in toxicity

and bullying directed at Asians. The study also revealed an increase in hate speech among children and teens that correlated directly with an increase of COVID-19 infections within the populace. According to Verywell, that increase was due in part to the extra leisure time and online presence that children had due to lockdown and online schooling. A report from Common Sense Media indicated that children and teens spent around 17 percent more time on social media sites due to the pandemic. Psychological reasons, including self-preservation and self-defense behaviors, have also been cited (by Verywell) as possible causes for the sudden rise in cyberbullying and online toxicity during the pandemic. There are a large number of ways parents can respond to cyberbullying, but it appears the most common response is to talk to children about online safety. Comparitech found 59.4% of parents talked to their children about internet safety and safe practices after cyberbullying occurred. Parents may need to take more steps to intervene, however, as only 43.4% identified adjusting parental controls to block offenders, only 33% implemented new rules for technology use, and only 40.6% saved the evidence for investigators. Very few parents (just 34.9%) notified their child's school about cyberbullying. And a small number (10.4%) took the nuclear option and completely took away their child's technology in response.

A 2022 Pew Research study found that nearly half of all teens (49%) had experienced some form of cyberbullying. The most common type was offensive name calling, but one in ten had also received physical threats. Another study from 2021 shows that this isn't unique to teens, with around 40 percent of Americans under 30 having experienced online harassment. Of these, 50% identified politics as the reason behind the incident. Among teens, the most common specific types of cyberbullying include:

- Offensive name-calling (32 percent)
- Spreading of false rumors (22 percent)
- Receiving unsolicited explicit images (17 percent)
- Repeated requests for their location or whereabouts (15 percent)
- Physical threats (10 percent)
- Having explicit images of them shared without their consent (7 percent)

A 2021 study from the UK anti-bullying organization Ditch the Label found that over 40 percent of people under 25 years old aren't sure whether social media platforms should be more tightly moderated. Around a third would like to see increased moderation, with 15 percent of respondents being against this move.

Cyberbullying and online gaming:

Social media tends to eat up most of the attention related to cyberbullying, but it can occur across any online medium, including online gaming. In one survey, 90 percent of gamers reported experiencing cyberbullying in-game, with racism, hate speech, and extremist content extremely common. Meanwhile, a survey of over 2,000 adolescents found that over one-third experienced bullying in mobile games. And a 2020 Ditch the Label survey of over 2,500 young adults found 53% reported to be victims of bullying in online gaming environments, while over 70% believe bullying in online games should be taken more seriously. Unfortunately, Ditch the Label's 2019 survey found that the number of respondents who had been bullied in an online game had risen to 76% (although confusingly, this figure dropped to just 11% in 2020 — the reasons why are unclear but should hopefully come to light with further research). Online gaming bullying can extend beyond just hurtful words. It can also include the dangerous activity known as swatting, in which perpetrators locate the home address of the victim and make a false criminal complaint to the victim's local police, who then "send in the SWAT team" as a response. Swatting has resulted in the shooting death of innocent victims, making it a particularly troubling practice more commonly associated with the gaming community.

Types of cyberbullying:

1. Exclusion

Exclusion is the act of leaving someone out deliberately. Exclusion exists with in-person bullying situations, but is also used online to target and bully a victim. For example, your child might be excluded/uninvited to groups or parties while they see other friends being included, or left out of message threads or conversations that involve mutual friends.

2. Harassment

Harassment is a broad category under which many types of cyberbullying fall into, but it generally refers to a sustained and constant pattern of hurtful or threatening online messages sent with the intention of doing harm to someone.

3. Outing/Doxing

Outing, also known as doxing, refers to the act of openly revealing sensitive or personal information about someone without their consent for purposes of embarrassing or humiliating them. This can range from the spreading of personal photos or documents of public figures to

sharing an individual's saved personal messages in an online private group. The key is the lack of consent from the victim.

4. Trickery

Trickery is similar to outing, with an added element of deception. In these situations, the bully will befriend their target and lull them into a false sense of security. Once the bully has gained their target's trust, they abuse that trust and share the victim's secrets and private information to a third party or multiple third parties.

5. Cyberstalking

Cyberstalking is a particularly serious form of cyberbullying that can extend to threats of physical harm to the child being targeted. It can include monitoring, false accusations, threats, and is often accompanied by offline stalking. It is a criminal offense and can result in a restraining order, probation, and even jail time for the perpetrator.

6. Fraping

Fraping is when a bully uses your child's social networking accounts to post inappropriate content with their name. It can be harmless when friends write funny posts on each other's profiles, but has the potential to be incredibly harmful. For example, a bully posting racial/homophobic slurs through someone else's profile to ruin their reputation.

7. Masquerading

Masquerading happens when a bully creates a made up profile or identity online with the sole purpose of cyberbullying someone. This could involve creating a fake email account, fake social media profile, and selecting a new identity and photos to fool the victim. In these cases, the bully tends to be someone the victim knows quite well.

8. Dissing

Dissing refers to the act of a bully spreading cruel information about their target through public posts or private messages to either ruin their reputation or relationships with other people. In these situations, the bully tends to have a personal relationship with the victim, either as an acquaintance or as a friend.

9. Trolling

Trolling is when a bully will seek out to intentionally upset others by posting inflammatory comments online. Trolling may not always be a form of cyberbullying, but it can be used as a tool to cyberbully when done with malicious and harmful intent. These bullies tend to be more detached from their victims, and do not have a personal relationship.

10. Flaming

This type of online bullying consists of posting about or directly sending insults and profanityto their target. Flaming is similar to trolling, but will usually be a more direct attack on a victim to incite them into online fights.

Consequences of cyberbullying:

Emotional Effects of Cyberbullying

Not surprisingly, cyberbullying is a significant stressor in a young person's life. In fact, research shows that 32% of kids who are targets of cyberbullying report experiencing at least one symptom of stress. In addition to feeling distressed, they also may feel embarrassed, hurt, and even fear for their safety. They may even blame themselves for the cyberbullying.

Humiliation

Because cyberbullying occurs in cyberspace, online bullying feels permanent. Kids know that once something is out there, it will always be out there. They can feel exposed, embarrassed, and overwhelmed. When cyberbullying occurs, nasty posts, messages, or texts can be shared with multitudes of people. The sheer volume of people that know about the bullying can lead to intense feelings of humiliation.

Isolation

Cyberbullying sometimes causes kids to be excluded and ostracized at school. Consequently, they often feel alone and isolated.1 This experience can be particularly painful because friends are crucial at this age. When kids don’t have friends, this can lead to more bullying. When cyberbullying occurs, parents sometimes recommend shutting off the computer or turning off the cell phone. But for many kids, using these devices is considered the most important way they communicate with others. Turning them off often means cutting off their connection with their world, which can make them feel more secluded.

Anger

Many victims of cyberbullying will get angry about what is happening to them. In fact, research indicates that anger is the most common response to cyberbullying (followed by being upset and worried).1 Some kids that are victimized may even plot revenge and engage in retaliation. Aside from the trouble they could get into, this approach is dangerous becauseit can keep them locked in the bully-victim cycle. While it's always better to forgive a bully than it is to get even, this is often easier said than done. If your child seems intensely angry over cyberbullying, it may help for them to speak with a counselor or therapist who can teach them to channel that anger in productive ways.

Powerlessness

Victims of cyberbullying often find it difficult to feel safe. They may feel vulnerable and powerless. Typically, these feelings surface because online bullying can invade their home through a computer or cell phone at any time of day. They no longer have a place where they can escape.

Mental Effects of Cyberbullying

When cyberbullying is ongoing, victims may relate to the world around them differently than others. For many, life can feel hopeless and meaningless. They may lose interest in things they once enjoyed and spend less time interacting with family and friends. And, in some cases, depression and thoughts of suicide can set in.

Depression and Anxiety

Victims of cyberbullying may succumb to anxiety, depression, and other stress-related conditions. The added stress of coping with cyberbullying on a regular basis can steal their feelings of happiness and contentment. It also can increase feelings of worry and isolation. Research has consistently supported the notion that increasing levels of cyberbullying lead to higher levels of depression. In fact, one study found that 93% of those victimized by cyberbullying reported feelings of sadness, powerlessness, and hopelessness.

Low Self-Esteem

Cyberbullying often zeros in on what already makes victims feel most vulnerable. Forexample, maybe a child who feels insecure about a birthmark ends up being bullied about just that. Even when that's not the case, though, online bullying can have an impact on self- esteem. Targets of bullying may begin to feel intense dissatisfaction with who they are.

As a result, they can begin to doubt their worth and value. Researchers speculate that because young people have an intense psychological need to be part of and accepted by a peer group, cyberbullying may cause psychological maladjustment, reduced well-being, and ultimately low self-esteem.

Academic Issues

Kids being victimized by cyberbullying may lose interest in school. As a result, they often have much higher rates of absenteeism than non-bullied kids. They may skip school to avoid facing the kids cyberbullying them or because they are embarrassed and humiliated by the messages that were shared online. Their grades may also suffer because they find it difficult to concentrate or study. And in some cases, kids may either drop out of school or lose interest in continuing their education after high school.

Suicidal Thoughts and Self-Harm

Sometimes targets of cyberbullying respond to their intense feelings by harming themselves in some way. For instance, some might engage in self-harm such as cutting or burning themselves. In fact, research has consistently linked bullying and self-harm. Cyberbullying also increases the risk of suicide. Kids that are constantly tormented by peers through text messages, instant messaging, social media, or apps often begin to feel hopeless and that the only way to relieve the pain is ending their life. As a result, they may fantasize about dying in order to escape.

Behavioral Effects of Cyberbullying

Kids who are cyberbullied may display the same behavioral changes as those who are bullied in more traditional ways. For example, they exhibit a loss of interest in activities and engage in secretive behavior. In extreme cases, or when cyberbullying is prolonged, kids sometimes even exhibit more significant behavioral changes.

- Using drugs or alcohol: Kids who are harassed online are more likely to engage in substance abuse. In fact, one study found that targets of cyberbullying were 2.5 times more likely to use marijuana or engage in binge drinking than their peers.
- Skipping school: Sometimes when kids are cyberbullied, the thought of going to school is just more than they can handle. Consequently, it's not uncommon for them to skip school or even behave in such a way that results in suspension. In one survey,

those who were cyberbullied reported two or more suspensions or detentions in the prior year.

- Carrying a weapon: Even more concerning is the fact that kids who are cyberbullied are more likely to bring a weapon to school. In fact, one survey found that targets of cyberbullying were eight times more likely to have brought a weapon to school in the last 30 days than their peers.

Physical Effects of Cyberbullying

Being targeted by cyberbullies can be crushing, especially if a lot of kids are participating in it. The feelings of overwhelm and stress can manifest physically, which issues such as:

- Gastrointestinal issues: The stress of bullying also can cause or worsen conditions like upset stomach, abdominal pain, and stomach ulcers. Kids may also struggle with frequent nausea, vomiting, and diarrhea.
- Disordered eating: Kids who are cyberbullied may experience changes in eating habits like skipping meals or binge eating. Because their lives feel out of control, theylook to their eating patterns as something they can control. These efforts may morph into a full-blown eating disorder, especially if the bullying has caused a distorted bodyimage.
- Sleep disturbances: Experiencing cyberbullying can impact a person's sleep patterns. They may suffer from sleep issues like insomnia, sleeping more than usual, or nightmares.

Cyber bullying in India:

Not only in India, people these days are getting bullied online all across the world. The worst part is there is no awareness regarding kids getting bullied by which they are getting mentally disturbed. Many videos of people getting bullied are uploaded on YouTube, and they cannot be traced because those videos were uploaded anonymously. In many countries there are no specific laws for cyberbullying.

United Kingdom: In the U.K half of 12-15 years kids get bullied each and every day. Though bullying is not a criminal offense in U.K there are many laws which can be used to punish a person who bullied someone such as Protection from Harassment Act,1997 where harassmentis punished under section 3, Computer Misuse Act, 1990, crime defamation Acts 1952 and 1996

United States: Nearly every state in the U.S took steps to prevent bullying or cyberbullying. A new law was passed to make cyberbullying a crime under the Megan Meier Cyber bullying prevention Act, 33. California passed an act Safe Place To Learn Act to make schools and colleges a better place to learn and the penalties are suspension, 1 year of jail and fine up to $1000.

Italy: In May, 2017 Italy passed a new law with 432 votes in which cyber bullying is described as an offense. This law is passed after many victims committed suicide and most ofthe victims were teenagers.

What Are The Steps Taken?

There are no specific laws in India regarding cyber bullying. Cyber bullying is not something which can be obliterated without even reporting. It is not easy for a victim to cope up with the bully they faced because someone said words that scare, rumors destroy and bullies kill. After the amendment of Indian Penal Code, 1860 in 2013 there are some laws to rely on such as Section 499 of IPC defines defamation, Section 292A defines printing matter intended to blackmail, Section 354A describes sexual harassment, Section 354D defines stalking, Section 509 defines any word or act intended to insult a woman. The Information Technology Amendment Act also provides remedies for cyber bullying. Section 66A of IT Act defines punishment for a person for sending an offensive message through any communicating device. Section 66E defines punishments for invading privacy. Section 67 defines punishment for publishing any obscene picture.

What Is The Present Condition? Presently in India there is a huge increase in cyberbullying cases. But the no.of cases reported are not proportionate to actual no.of cases because 9.2% of the kids didn't tell their teachers and parents about getting bullied. According to Child Rights and You (CRY) 1 in 3 adults get bullied everyday and most of them are aged between 13-18 years. According to the National Crime Records Bureau there is a 36% increase in cyber stalking and cyber bullying cases in India.

Rittika Sharma's case: Rittika Sharma, who was a student in a reputed Delhi school was stalked by a Facebook friend whom she unfriended months ago and whom she gave all her information including residential address, school address and even cell phone number. She told her brother regarding this and her brother filed a complaint against this. After this incident Delhi police organized an awareness program where all the students were told not to send their personal details to any stranger.

Ritu Kohli's Case: While discussing cyber stalking and Cyber bullying, Ritu Kohli's case is the case one should mention. Ritu Kohli's Case was the first cyber stalking case reported in India. A girl named Ritu Kohli filed a complaint in 2001 that someone else is using her identity in social media and she was deliberately getting calls from different numbers. She was also getting calls from abroad. A case was also filed under Section 509 of Indian penal code.

The role of social media in cyberbullying

As people become more familiarized with and exposed to social media, the opportunity to cyberbully increases. Social media platforms that allow free and open commenting can become a very fearful environment for cyberbullying victims, where threats, aggressive, demotivating, or offensive comments or messages, or edited pictures or videos, can be made and shared outside of the victim's control before they have a chance to respond. Embarrassment over the issue can lead to people hiding online bullying from their friends andfamily in real life, further fuelling feelings of isolation, depression, and anxiety. A lack of awareness and support can also create a barrier for the victims to open up about their problems and lead to unstable mental health.

Twitter Cyberbullying:

Cyberbullying on Twitter refers to using Twitter to send or post messages that might be considered offensive, hurtful, or mean. Cyberbullying on Twitter is a serious problem becauseit can have long-lasting effects on victims.

Twitter has recently announced that it will be rolling out a new “quality filter” that is designed to “remove all Tweets from your notification timeline that contain threats, offensiveor abusive language, duplicate content, or are sent from suspicious accounts.” The “quality filter” is only attached to verified users since they have the most followers and therefore are susceptible to the most abuse, but Twitter has also implemented other anti-harassment tools such as a feature that makes it easier to report abuse to law enforcement. So essentially, this quality filter and other recent features are designed to prevent instances of cyber-bullying and protect user safety. Cyber-bullying is more and more common as Internet users are shielded by anonymity on the Web. Cyber-bullying is especially present on Twitter. According to data from the Pew Center, Twitter users face many forms of harassment including death threats and threats of sexual abuse and stalking and the victims of this abuse are disproportionately

women. There have been several recent high-profile cases of cyber-bullying involving Twitter including #gamergate, the harassment of Robin William's daughter after his death, and Ashley Judd's decision to press charges against trolls. These high-profile incidents have been speculatively identified as the impetus for Twitter's implementation of anti-harassment blocking tools including the "quality filter". Twitter initially positioned itself as the "free speech wing of the free speech party", which meant that they took a neutral view on message content. Twitter's "neutral view" has seemingly made the company more tolerant of abuse and harassment on their social media site relative to other social media sites. For instance, Twitter is notoriously criticized for its failure to deal with cyber-bullying. In fact, Twitter's CEO Dick Costolo claimed that " We [Twitter] suck[s] at dealing with abuse", apologized for his company's failure to adequately protect its users from abuse via Twitter, and admitted that cyber-bullying has cost platform users. The "quality filter" and other blocking tools have emerged since Twitter's CEO has taken personal responsibility for Twitter's slow response to protecting its users. Twitter has no legal obligation to censor its users but Twitter is also not under any obligation imposed by the First Amendment to protect free speech. Therefore, as a private company, Twitter may balance free speech against user safety in any manner it so chooses. Given the bad rap Twitter is receiving for not censoring enough and the resulting loss of platform users both low and high-profile, it is likely a good decision for Twitter to implement more anti-harassment blocking tools. Free speech is an admirable value but it likely shouldn't come at such a high cost to user safety.

The most common form of cyberbullying on Twitter is sending out tweets with offensive or insulting content to the victim's followers. The tweets are usually sent from an account with very few followers and follow only one person – the victim.

This type of bullying is called "flaming." Flamers often send out insults and derogatory statements to their victims to provoke a response.

Twitter Cyberbullying Policy

Twitter has been criticized for its inability to control cyberbullying on its platform. For this reason, Twitter has recently updated its policy to address these concerns. The new policy will prohibit behaviors such as "hateful conduct" and "specific threats of violence or wishing for serious physical harm." The company will also be taking steps to protect victims by allowing them to report abuse directly from their account settings page.

Cyberbullying on Twitter Messages Examples

Cyberbullying can take many forms and can happen at any time. It can occur through email, text messaging, instant messaging, social networking sites such as Facebook and Instagram, or other online forums where people share their thoughts and ideas.

Cyberbullying on Twitter Statistics

Recent surveys by the authorities found that 7200 US teenagers use social media. The results indicated that 59% of teenagers surveyed used Twitter and social networking sites such as Facebook to communicate with one another.

The world has led to increased cyberbullying, especially among teenagers. Researchers have widely reported the issues surrounding teenagers and cyberbullying.

Cyberbullying uses information and communication technology to harass and harm deliberately, repetitively, and hostilely. Ditch the Label surveyed more than 10,008 teenagers and discovered that of the 43% who used Twitter, 28% had experienced cyberbullying.

How Twitter Is Fighting Harassment & Cyberbullying

1. Expanded notification filtering

TwitterTwitter users can use this tool to filter which types of accounts they receive notifications from. For example, if you don't want to receive notifications from a user withouta profile photo, you could specify that. This tool is meant to filter out abuse from unverified accounts or specific people users have identified as unwanted.

1. More ways to mute content

Twitter expanded on the mute button's capabilities so users can mute keywords or entire phrases from their notifications sections. Users can also decide how long they want to mute those words -- whether it be for a day, a month, or indefinitely. In this way, you can customize which content you see in your notifications and when you see it.

2. Greater transparency around reporting.

Whereas previously, users had a hard time understanding when or if their reports of abuse were even being processed, Twitter is now providing transparency. Users will receive notifications when and if Twitter decides to take action so they can keep track of previous reporting.

3. Twitter "time-out"

In a recent article, BuzzFeed reported that some Twitter users were seeing another new feature, similar to the time-out we all experienced as children (unless you were better behaved than I was). If users' tweets are flagged as abusive or otherwise in violation of Twitter Rules, their tweets are temporarily limited from view by users who don't follow them. Hopefully neither you nor your brand's Twitter will see this notification, but the company is hoping it will send a message to abusers to stop what they're tweeting or risk furtherpunishment.

4. Safer search results.

Machine-learning algorithms will filter search results so users aren't served content from accounts that have been reported, muted, or otherwise marked as abusive. The content will still be on Twitter if users are really looking for it, but if it could potentially be abusive, it won't be served up as a primary search result.

5. Collapsing abusive tweets.

Twitter will start identifying and hiding tweets that are deemed "low quality" or from potentially abusive accounts so users see the most relevant conversations first. Like the safe search feature, those tweets will still be on Twitter -- but users have to search for them specifically.

6. Stopping creation of new abusive accounts.

Using another algorithm, Twitter will prevent abusive and flagged users from creating multiple new accounts they can use to spam and harass other users. The algorithm will scan for multiple accounts from the same email addresses and phone numbers, for example, as a way to spot potential bullies.

Chapter 2

LITERATURE SURVEY

The literature survey has been conducted on different papers such as technical papers and review papers in the domain published in leading publications, journals and conferences. Keywords suchas cyber bullying detection, twitter cyberbullying, machine learning cyberbullying, NLP cyberbullying, etc are used to filter out.R. R. Dalvi et al. concentrates on basic conventional machine learning models like Support Vector Machine (SVM) and Naive Bayes algorithm in the cyberbullying detection task. They used two classifiers i.e. SVM and Naïve Bayes are used for training and testing the social media bullying content. Through extensive experimentation, we assess the effectiveness of both classifiers in accurately identifying instances of bullying within the dataset. However, it is noteworthy that SVM outperforms Naïve Bayes in this context, demonstrating superior performance in comparison to similar work conducted on the same dataset. This highlights the effectiveness of SVM in accurately identifying instances of bullying within social media content. But SVM outperforms Naive Bayes of similar work on the same dataset. Their methodology incorporates the utilization of the Twitter API to collect tweets, which are subsequently passed through the trained machine learning model to determine whether they contain instances of bullying. This integration of the Twitter API streamlines the process of data collection and enables real-time detection of bullying behavior within the platform [1]. T. Balet etal. have delved into the utilization of various machine learning models, including Logistic Regression (LR), k-Nearest Neighbors (KNN), Random Forest (RF), Classification and Regression Trees (CART), Naive Bayes, and Support Vector Machine (SVM). The objective has been to train and assess the performance of these models using datasets comprising Twitter tweets, thereby enabling the identification of cyberbullying instances within social media discourse. The methodologies employed encompassed rigorous training and testing processes, wherein each algorithm underwent evaluation based on diverse performance metrics. These metrics included accuracy, F1-score, cross-validation score, and receiver operating characteristic (ROC) curve analysis. Through meticulous examination, researchers sought to ascertain the efficacy of each algorithm in accurately detecting and classifying instances of abusive or harassing language within the Twitter dataset. The results of these investigations unveiled intriguing findings regarding the performance of the machine learning models. Notably, the Random Forest algorithm emerged as the frontrunner, demonstrating exceptional accuracy with a notable 94% success rate. This outcome underscored the robustness and effectiveness of Random Forest in discerning instances of cyberbullying within social media

communications. However, it is imperative to acknowledge the performance variations among the different algorithms, eachpresenting unique strengths and limitations in the context of cyberbullying detection [2].

S. A. Mathur et al., in their literature, a real-time cyber-bullying detection system for Twitter utilizing Natural Language Processing (NLP) and Machine Learning (ML) has been proposed. They have endeavored to train the system on a dataset comprising cyberbullying tweets, employing several ML algorithms for analysis and comparison of their performance. Notably, the study found that Random Forest emerged as the most effective algorithm following rigorous tuning processes. The emphasis on achieving real-time analysis prompted the utilization of Selenium for tweet scraping purposes, enabling the retrieval of tweets from designated Twitter accounts while storing their corresponding timestamps for efficient tracking of already processed tweets. Additionally, an image captioning model was integrated into the system architecture, facilitating the generation of descriptions for images posted on Twitter accounts. These descriptions were subsequently compared with user-written captions to identify and filter out potential spam tweets. Through meticulous experimentation and analysis, the study has underscored the significant impact of ML algorithms' selection and preprocessing techniques on cyberbullying detection efficacy within the Twitter domain. Notably, the findings elucidate the pivotal role of appropriate algorithm selection and preprocessing methodologies in enhancing the performance of cyberbullying detection mechanisms. The outcomes of this study illuminate the efficacy of ML algorithms in combating cyberbullying, offering valuable insights into their appropriateness and effectiveness for detection purposes. By shedding light on the intricacies of ML algorithm selection and preprocessing techniques, the research contributes to the broader discourse surrounding cyberbullying prevention strategies within online platforms such as Twitter. Furthermore, the integration of real-time analysis capabilities and advanced image captioning models underscores the multifaceted approach adopted by researchers to address the complex challenges associated with cyberbullying detection in contemporary digital landscapes [3]. Salawu et al. have introduced a novel English Twitter-based dataset tailored for cyberbullying detection and the identification of online abuse. This dataset, encompassing a substantial volume of 62,587 tweets, was curated from Twitter utilizing specific query terms meticulously designed to retrieve tweets characterized by varying forms of bullying and offensivecontent. These forms include insult, trolling, profanity, sarcasm, threat, pornographic material, and exclusionary language. The methodology involved the recruitment of a diverse pool comprising 17 annotators tasked with conducting fine-grained annotation on the dataset, with each tweet being meticulously annotated by three annotators. It is noteworthy that all annotators

possess high school education credentials and are frequent users of social media platforms. Inter-rater agreement, a crucial metric for assessing dataset reliability and consistency, was evaluated utilizing Krippendorff's Alpha, yielding a commendable coefficient of 0.67. This metric underscores the robustness and reliability of the annotated dataset, thereby enhancing its utility for subsequent analysis and experimentation. Notably, comprehensive analysis conducted on the dataset corroborated the prevalence of common cyberbullying themes reported by earlier studies, while also uncovering intriguing relationships between distinct classes of cyberbullying content. Subsequently, the curated dataset served as the foundational training corpus for a range of transformer-based deep learning models. These models, leveraging the transformative capabilities of deep learning architectures, exhibited impressive performance in cyberbullying detection tasks. The utilization of transformer-based models underscores the adoption of state-of- the-art methodologies to address the inherent complexities and nuances associated with cyberbullying detection within online platforms such as Twitter [4].

Muneer A et al. have introduced an ensemble stacking learning approach for detecting cyberbullying on Twitter, employing a combination of Deep Neural Network (DNN) methods. Additionally, they have introduced BERT-M, a modified variant of the BERT model, to further enhance cyberbullying detection capabilities. The dataset utilized in this study was sourced from Twitter and subjected to preprocessing to eliminate irrelevant information and enhance dataset quality. The feature extraction process entailed leveraging word2vec with Continuous Bag of Words (CBOW) to generate weights in the embedding layer. These extracted features were subsequently processed through a convolutional and pooling mechanism, aimed at reducing dimensionality while capturing position-invariant characteristics of offensive language. The validation of the proposed stacked model and BERT-M was conducted using established model evaluation measures. Notably, the stacked model exhibited impressive performance metrics, achieving an F1-score of 0.964, precision of 0.950, and recall of 0.92. Furthermore, the detection time for the stacked model was reported at 3 minutes, surpassing previously reported accuracy and speed scores for all known Natural Language Processing (NLP) detectors of cyberbullying, including standard BERT and BERT-M models. The experimental results revealed that the stacking ensemble learning approach achieved exceptional accuracy rates of 97.4% and 90.97% in detecting cyberbullying on Twitter and combined Twitter-Facebook datasets, respectively. These findings underscore the efficacy of the proposed ensemble learning approach in detecting cyberbullying across social media platforms (SM), emphasizing the significance of combining multiple models to enhance performance [5].

Fati SM et al. have undertaken a comparative analysis of deep learning methodologies to assess their efficacy in addressing cyberbullying within a globally recognized Twitter dataset. To confront the challenges inherent in identifying abusive tweets, attention-based deep learning techniques have been introduced. These methodologies leverage word2vec with Continuous Bag of Words (CBOW) to construct embedding layer weights, facilitating feature extraction. The resulting feature vector is then processed through a convolutional and pooling mechanism, serving to reduce feature dimensionality while capturing position-invariant characteristics of offensive language. Feature classification is subsequently performed using a SoftMax function. Upon thorough evaluation using benchmark experimental datasets and established evaluation metrics, the convolutional neural network model incorporating attention-based long- and short- term memory mechanisms emerged as the most effective deep learning approach for cyberbullying detection. Notably, this model outperformed other deep learning methodologies examined in the study, demonstrating superior performance in identifying and classifying abusive content within Twitter datasets. The proposed cyberbullying detection methods were rigorously evaluated using benchmark experimental datasets and well-established evaluation metrics, culminating in the identification of the attention-based 1D convolutional long short-term memory(Conv1DLSTM) classifier as the most proficient among the implemented methodologies. The results obtained underscore the effectiveness of attention-based deep learning techniques, particularly the Conv1DLSTM classifier, in addressing the pervasive issue of cyberbullying within social media platforms such as Twitter [6]. Parikh, R et al. have has proposed a compelling approach for detecting instances of cyberbullying on Twitter through the utilization oftransformer architectures. Four distinct transformer models were trained and evaluated, with theirperformance compared against each other. Notably, BERT-Base-Uncased emerged as the top- performing transformer architecture, achieving a commendable test accuracy of 85.81% and an F1-score of 0.8566. This performance surpassed that of other transformer models, including DistilBERT-Base-Uncased, ELECTRA, and MobileBERT-Uncased. Furthermore, comparative analysis with traditional machine learning algorithms revealed that BERT-Base-Uncased consistently produced superior results, thereby affirming its effectiveness for real-time identification of cyberbullying instances. The findings underscore the efficacy of transformer architectures, particularly BERT-Base-Uncased, in detecting and mitigating instances of cyberbullying within the Twitter platform. Through rigorous evaluation and comparison, researchers have demonstrated the superior performance of BERT-Base-Uncased in accurately identifying malicious content, thus highlighting its potential as a robust tool for combating

cyberbullying in real-time scenarios [7]. Raj M et al. introduces a deep learning framework aimedat evaluating real-time Twitter tweets or social media posts to accurately identify cyberbullying content. Recent research indicates that deep neural network-based approaches exhibit greater efficacy in detecting cyberbullying texts compared to traditional techniques. Furthermore, the application developed within this study demonstrates proficiency in recognizing cyberbullying posts composed in English, Hindi, and Hinglish, thereby encompassing multilingual data. The primary objective of this research endeavor is to construct a CNN-BiLSTM deep learning detection model capable of identifying cyberbullying content within tweets across three distinct languages in real-time datasets. This model leverages Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) architectures to effectively discern and classify cyberbullying instances. Additionally, the study culminates in the development of a website resembling a social media platform, serving as a tangible manifestation of the proposed model's applicability and functionality. The literature review highlights the innovative contributions made by researchers in the realm of cyberbullying detection, particularly through the utilization of deep learning methodologies. By introducing a CNN-BiLSTM deep learning framework, researchers have demonstrated the potential to accurately identify cyberbullying content in real-time social media data across multiple languages. The application's proficiency in detecting cyberbullying instances in English, Hindi, and Hinglish underscores its versatility and applicability across diverse linguistic contexts [8]. B. A. H. Murshed, et al. propose a hybrid deep learning model, DEA-RNN, aimed at combating the increasingly prevalent issue of cyberbullying (CB) on social media platforms, particularly Twitter. In response to the urgent need for safer social media environments, this paper introduces a novel approach that merges Elman type Recurrent Neural Networks (RNN) with an optimized Dolphin Echolocation Algorithm (DEA). The combination of these techniques serves to fine-tune the parameters of the Elman RNN and reduce training time, enhancing the efficiency of CB detection. To thoroughly evaluate the proposed DEA-RNN model, the authors conducted extensive experiments using a dataset comprising 10,000 tweets. They compared the performance of DEA-RNN against state-of-the-art algorithms including Bi-directional Long Short-Term Memory (Bi-LSTM), SVM, Multinomial Naive Bayes (MNB), and Random Forests (RF). Remarkably, the experimental results consistently demonstrated the superior performance of DEA-RNN across various scenarios, showcasing its effectiveness in detecting CB on the Twitter platform. In scenario 3, DEA-RNN exhibited exceptional performance metrics, achieving an average accuracy of 90.45%, precision of 89.52%, recall of 88.98%, F1-score of 89.25%, and specificity of 90.94%. These results highlight the potential of the proposed hybrid deep learning model to significantly outperform

existing approaches in the critical task of cyberbullying detection on social media networks like Twitter [9]. Hasan MT et al. Hasan MT et al. provide a comprehensive literature review aimed at identifying the gaps in existing surveys and presenting a deep-learning-based defense ecosystem for cyberbullying detection. The paper delves into data representation techniques and various deep-learning-based models and frameworks utilized in cyberbullying detection. Through critical analysis, the authors assess the contributions of existing deep learning (DL) techniques and outline future research directions. The review highlights the datasets commonly used in DL- based cyberbullying detection, along with the specific DL architectures employed and the tasks accomplished for each dataset. By summarizing the key findings from previous studies, the authors offer insights into the effectiveness of DL approaches in addressing cyberbullying on social media platforms. Furthermore, the paper elucidates the challenges faced by researchers in this domain and discusses open issues that warrant further investigation. By addressing these challenges and unresolved issues, future research endeavors can strive to enhance the efficacy and robustness of DL-based cyberbullying detection systems, thereby contributing to the development of safer online environments [10].

Obamiyi, S et al. introduce an ensemble model for cyberbullying detection, a significant measure in combating online harassment. The model employs a majority voting ensemble approach, integrating three supervised machine learning classifiers: Support Vector Machine (SVM), Naive Bayes (NB), and K-Nearest Neighbors (K-NN). Utilizing a dataset of malignant comments sourced from Kaggle.com, the authors split the data into training and evaluation sets at a ratio of 70:30 to optimize model performance. Evaluation metrics were applied to assess the model's effectiveness, with the ensemble model demonstrating superior performance compared to individual classifiers, achieving an impressive accuracy of 95%. Notably, the ensemble model exhibited consistency across various metrics, underscoring its efficacy in detecting cyberbullying comments effectively. This research underscores the importance of ensemble learning techniques in enhancing the robustness and accuracy of cyberbullying detection systems, thereby contributing to the creation of safer online environments [11]. Nitya Harshitha, T. et al. present a novel approach for text classification using a hybrid random forest-based Convolutional Neural Network (CNN) model, leveraging the advantages of both methodologies. The study involved the collection and annotation of real-time datasets from Twitter and Instagram, demonstrating the efficacy of the proposed technique. Through comparative analysis, the performance of various Machine Learning (ML) and Deep Learning (DL) algorithms was evaluated, with the hybrid RF- based CNN model surpassing them in terms of accuracy and execution speed. This is

particularly significant in the context of timely detection of bullying incidents and providing support to affected individuals. Notably, the hybrid model achieved an impressive accuracy rateof 96% and exhibited a 3.4-second improvement in execution speed compared to conventional CNN models. The findings highlight the potential of hybrid ML-DL approaches in enhancing text classification tasks, thereby contributing to more effective and efficient cyberbullying detection systems [12]. Selvakumar, M. et al. address the limitations of traditional techniques in effectively detecting user interests and active communities. To overcome these challenges, they propose the utilization of a Densenet Convolution Neural Network (DnetCNN) approach for community detection. The study begins by collecting a dataset from the Kaggle repository and preprocessing it to eliminate inconsistencies and missing values. Additionally, the authors introduce the User Behavior Impact Rate (UBIR) technique to identify user URL access, key terms, and page access patterns. Furthermore, they employ the Web Crawling Prone Factor Rate (WCPFR) technique to detect malicious activities using random forest and decision methods. Subsequently, the Spider Web Cluster Community based Feature Selection (SWC2FS) algorithm is employed to select the most relevant attributes from the dataset. Leveraging these attributes, the DnetCNN approach is applied to identify community groups. The experimental results demonstrate the superior performance of the proposed method compared to alternative approaches. This research contributes to advancing community detection methodologies, particularly in the context of identifying user interests and active communities, offering promising implications for various domains reliant on community analysis [13].

Chapter 3
EXISTING SYSTEM

The twitter cyberbullying detection is basically a text classification problem. The input text string to the detection system will be classified into normal or cyberbullying. Or sometimes there will be multiple classes like different types of cyberbullying. Basically, it is a text classification problem. So, any classification algorithms such as conventional statistical algorithms or today's machine learning algorithms can be used for this twitter tweet text classification. Techniques in NLP (Natural Language Processing) can also be used.

RULE-BASED SYSTEMS

Traditional rule-based systems for Twitter cyberbullying detection operate on predefined sets of rules and patterns established to flag potentially abusive or harmful content. These rules are often crafted manually based on expert knowledge of cyberbullying behaviors and linguistic analyses. The system's effectiveness largely depends on the comprehensiveness andaccuracy of these rules. Rule-based systems typically employ a combination of keyword matching, pattern recognition, and linguistic analysis to identify cyberbullying behavior. For example, a rule may specify that any tweet containing explicit profanity or derogatory language is flagged as potentially abusive. Similarly, patterns of repeated harassment, such as targeted attacks against specific individuals or groups, can be identified through rule-based analysis. While rule-based systems offer simplicity and transparency in their approach, they are inherently limited by the static nature of their rules. Cyberbullying behaviors can be complex and constantly evolving, making it challenging for rule-based systems to keep pace with emerging threats. Additionally, rule-based systems may struggle to differentiate between genuine expressions of opinion or criticism and abusive behavior, leading to false positives or missed detections. Despite these limitations, rule-based systems can still play a valuable role in complementing more advanced detection techniques. By providing a baseline level of protection against known cyberbullying behaviors, rule-based systems can help reduce the volume of abusive content that needs to be processed by more resource-intensive methods, such as machine learning algorithms.

KEYWORD FILTERING

Keyword filtering is a straightforward approach to Twitter cyberbullying detection that involves monitoring Twitter streams for specific keywords or phrases commonly associated with abusive or harmful content. By identifying tweets containing these keywords, the system

can flag potential instances of cyberbullying for further analysis or intervention. Keyword filtering systems typically maintain a predefined list of keywords or phrases that are indicative of cyberbullying behavior. These keywords may include derogatory terms, explicit language, or threats of violence. The system continuously scans incoming tweets and compares them against the list of keywords to identify matches. One of the primary advantages of keyword filtering is its simplicity and computational efficiency. Unlike more complex machine learning algorithms, keyword filtering does not require extensive training or data processing. Instead, it relies on a straightforward matching process that can be implemented with minimal computational resources. However, keyword filtering also has significant limitations. One of the main challenges is the potential for high false positive rates. Since keywords alone may not capture the nuanced context of cyberbullying behavior, tweets containing flagged keywords may include legitimate expressions of opinion or criticism. As a result, keyword filtering systems must strike a balance between sensitivity andspecificity to minimize false positives while still detecting genuine instances of cyberbullying. Additionally, keyword filtering is inherently limited by the specificity of the keywords used. Cyberbullying behaviors can manifest in various forms and contexts, making it difficult to capture all instances with a predefined set of keywords. As a result, keyword filtering systems may fail to detect emerging or context-specific forms of cyberbullying that do not match the predefined keywords. Despite these limitations, keyword filtering can still serve as a valuable component of a broader cyberbullying detection strategy. When combinedwith other detection techniques, such as machine learning algorithms or social network analysis, keyword filtering can help identify potential instances of cyberbullying more effectively and efficiently.

SENTIMENT ANALYSIS

Sentiment analysis is a widely used technique in the field of natural language processing (NLP) that aims to discern the underlying sentiment expressed in text data. In the context of Twitter cyberbullying detection, sentiment analysis plays a crucial role in identifying tweets containing negative sentiment, which may indicate instances of abusive or harmful behavior. At its core, sentiment analysis involves analyzing the textual content of tweets to classify them as positive, neutral, or negative based on the sentiment conveyed. This classification is typically achieved using machine learning algorithms trained on labeled datasets, where each tweet is annotated

with its corresponding sentiment label. These algorithms learn to recognize patterns and linguistic cues associated with different sentiments, enabling them to accurately classify unseen tweets. In the context of cyberbullying detection on Twitter, sentiment analysis can help identify tweets containing negative sentiment that may be indicative of cyberbullying behavior. For example, tweets containing derogatory language, threats, or insults are likely to be classified as having negative sentiment. By focusing on tweets with negative sentiment, analysts can prioritize the detection and mitigation of potentially abusive content. However, it's important to note that sentiment analysis alone may not be sufficient for detecting cyberbullying behavior accurately. Cyberbullying can manifest in various forms and contexts, and not all instances may exhibit overtly negative sentiment. For example, subtle forms of cyberbullying, such as microaggressions or passive-aggressive comments, may not be adequately captured by sentiment analysis alone. By leveraging sentiment analysis alongside other detection methods, such as keyword filtering or social network analysis, analysts can gain a more comprehensive understanding of the presence and prevalence of cyberbullying behavior on Twitter. Additionally, ongoing advancements in sentiment analysis techniques, such as the incorporation of contextual information and deep learning models, hold promise for further improving its effectiveness in detecting cyberbullying behavior.

SOCIAL NETWORK ANALYSIS

Social network analysis (SNA) is a powerful methodology used to examine the structure and dynamics of social networks, including those formed on Twitter, to uncover patterns of interaction and identify key actors or communities within the network. In the context of Twitter cyberbullying detection, SNA offers valuable insights into the relationships and behaviors of users, allowing analysts to identify clusters or communities engaged in harassing behavior. At its core, social network analysis involves the study of nodes (representing individual users or entities) and edges (representing connections or interactions) within a network. By analyzing the topology of the network and the flow of information between nodes, analysts can identify influential users, detect communities or clusters, and uncover patterns of behavior indicative of cyberbullying. One common approach to social network analysis on Twitter is to construct a network graph where nodes represent Twitter users and edges represent interactions such as mentions, retweets, replies, and follows. By analyzing the structure of this graph, analysts can identify users who are central to the network (i.e., those with many connections) as well as communities or clusters of users who frequently interact with each other. In the context of cyberbullying detection, social network analysis can help identify patterns of interaction

indicative of abusive behavior. For example, clusters of users who frequently mention or target the same individual with harassing or derogatory comments may be flagged as potential cyberbullying perpetrators. Similarly, users who act as bridges between different communities or who have a disproportionately high number of followers may exert influence over the spread of abusive content within the network.

HUMAN MODERATION

Human moderation, often considered the gold standard in content moderation, involves the manual review and evaluation of user-generated content by human moderators to identify and address instances of cyberbullying on Twitter. This approach relies on human judgment and expertise to interpret the context and intent behind tweets, making it particularly effective in identifying nuanced forms of cyberbullying that may be challenging for automated systems to detect. At its core, human moderation involves a team of trained moderators who review incoming tweets flagged as potentially abusive or harmful. These moderators are typically equipped with guidelines or policies outlining acceptable and unacceptable behavior on the platform, providing them with a framework for making consistent and informed decisions. Using their judgment and expertise, moderators assess the content of flagged tweets, consider the surrounding context, and determine whether the tweets violate the platform's community guidelines or terms of service. One of the primary advantages of human moderation is its ability to capture the subtleties and nuances of cyberbullying behavior that may elude automated detection systems. Unlike machine learning algorithms, which rely on predefined patterns and features to identify abusive content, human moderators can interpret the intent behind tweets and consider the broader context in which they were posted. This enables them to identify and address emerging or context-specific forms of cyberbullying that may not be adequately captured by automated systems. Additionally, human moderation is inherently subjective and prone to bias, as moderators' judgments may be influenced by their own personal beliefs, experiences, and cultural backgrounds. This can result in inconsistencies in content moderation decisions and raise concerns about fairness and equity in the enforcement of platform policies. By complementing automated detection systems with human judgment and expertise, platforms can achieve a more comprehensive and effective approach to combating cyberbullying and fostering a safer and more inclusive online environment. Moreover, ongoing advancements in technology, such as the integration of artificial intelligence and natural language processing, hold promise for enhancing the efficiency and effectiveness of human moderation processes, enabling platforms to better

address the complex and evolving challenges of cyberbullying on Twitter and other social media platforms.

HYBRID APPROACHES

Hybrid approaches represent a sophisticated and multifaceted strategy that amalgamates various detection techniques, methodologies, and algorithms to create robust and effective detection systems. The fundamental principle behind hybrid approaches is the amalgamation of disparate methodologies and technologies to overcome the limitations inherent in individual techniques and achieve superior performance in identifying cyberbullying behavior on the platform. Hybrid approaches exhibit a dynamic interplay between different detection methods, seamlessly integrating their unique strengths to enhance the overall efficacy of the cyberbullying detection system. This integration fosters synergy among diverse detection techniques, thereby maximizingtheir collective impact in identifying and mitigating instances of cyberbullying on Twitter. One prevalent hybrid approach entails the integration of sentiment analysis with machine learning algorithms. Sentiment analysis serves as a preliminary filtering mechanism, discerning tweets with negative sentiment that may indicate potential instances of cyberbullying. Thesesentiment- labeled tweets are then subjected to machine learning algorithms for further analysis, leveraging the power of pattern recognition and classification to differentiate between cyberbullying and non-cyberbullying content. This synergistic fusion of sentiment analysis and machine learning enables the detection system to sift through vast volumes of Twitter data efficiently while enhancing the accuracy of cyberbullying detection. By scrutinizing the network topology and dynamics of Twitter interactions, social network analysis offers valuable insights into the underlying structure of the Twitter ecosystem, enabling analysts to pinpoint influential users, detect coordinated harassment campaigns, and identify vulnerable targets. This holistic understanding of the social network landscape augments the efficacy of cyberbullying detection efforts, empowering moderators to intervene proactively and disrupt the spread of abusive content within the Twitter community.Another facet of hybrid approaches involves the integration of human moderation into the detection pipeline, harnessing the nuanced judgment and contextual understanding of human moderators to discern subtle forms of cyberbullying that may elude automated algorithms. Human moderators play a pivotal role in evaluating the intent and context of flagged tweets, interpreting the nuances of language and social dynamics to make informed decisions regarding content moderation. This human-in-the-loop approach enhances the adaptability and flexibility of the detection system, enabling it to address emerging trends and evolving manifestations of cyberbullying effectively.One of the key strengths of hybrid

approaches lies in their adaptability to diverse use cases and environments. These approaches canbe tailored to accommodate specific requirements, constraints, and objectives, enabling analysts to customize the detection pipeline according to the unique characteristics of the Twitter data and the prevailing cyberbullying landscape. This versatility empowers detection systems to achieve optimal performance across varied contexts, ensuring robust and reliable cyberbullying detection capabilities. With ongoing advancements in technology and methodology, hybrid approaches are poised to play a pivotal role in mitigating the scourge of cyberbullying and promoting positive social interactions on social media platforms.

Chapter 4

PROPOSED SYSTEM

The proposed system is a machine learning-based system that uses a dataset after NLP-based preprocessing. The data as shown in the below diagram is collected and passed through as a series of NLP-based preprocessing and cleaning stages. The dataset contains a number of cyberbullying types and after cleaning and preprocessing, it is then split into Train and Test sets. Three models are trained on the prepared data. The best performing model is saved to be used in the backend of the twitter clone.

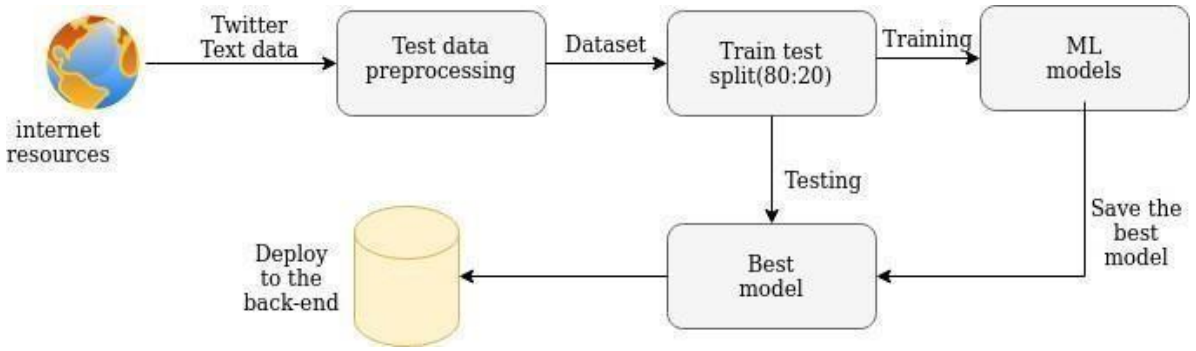


Figure4.1: The proposed system DFD

The twitter cyberbullying data is a labeled data collected from www.kaggle.com that has various cyberbullying types and labels in it. This raw tweet text will be preprocessed to get a well cleaned data suitable for machine learning model training. The preprocessing steps include basic statistical analysis, text lower case conversion, removing special characters, removing stop words and short forms, remove links and URLs, remove accent, normalize space, label encoding, Exploratory Data Analysis (EDA), Principal Component Analysis (PCA) and Vectorization.

Once all these preprocessing techniques have been applied on the data, it will be split into two; The Training set and the test set at 80:20 ratio. The former will be used for training the model and the later will be used for testing the trained models' performance. We go for developing the Machine Learning models with the help of python libraries and these models will be trained. After the testing and evaluation, we can confirm the best model and this model will be saved for deploying at the backend of the web app twitter clone.

The use case diagram below explains that the user can login to the system and type the tweet

he/she wants to tweet. Then once the user presses the “post” button, the text string will be sent to the backend to the trained model. It will generate the output based on the text as a cyberbullying tweet or not. The result will be sent to the front-end and the alert that the tweet cannot be posted will be displayed to the user. The tweet will be posted if the result is negative.

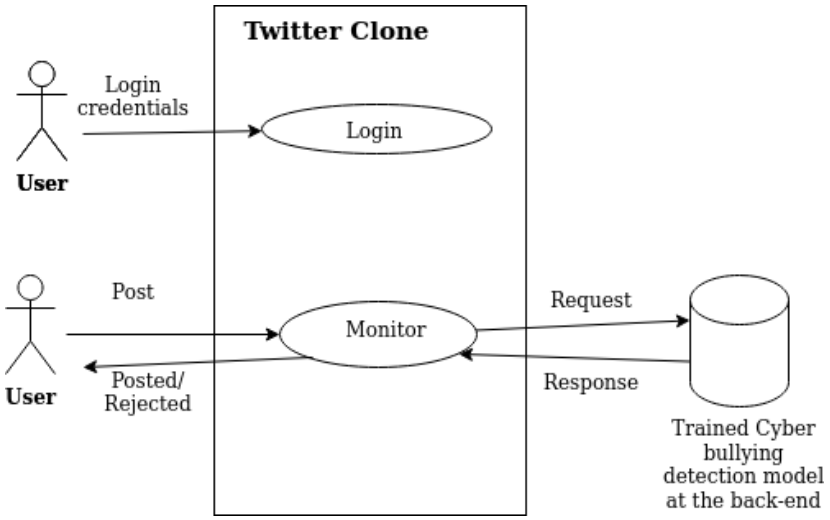


Figure 4.2: The use case diagram

Chapter 5

SYSTEM DESIGN AND ARCHITECTURE

The system architecture in the below figure shows the detailed explanation of the processingof the project right fromthe data collection and preprocessing to the final web app inference.

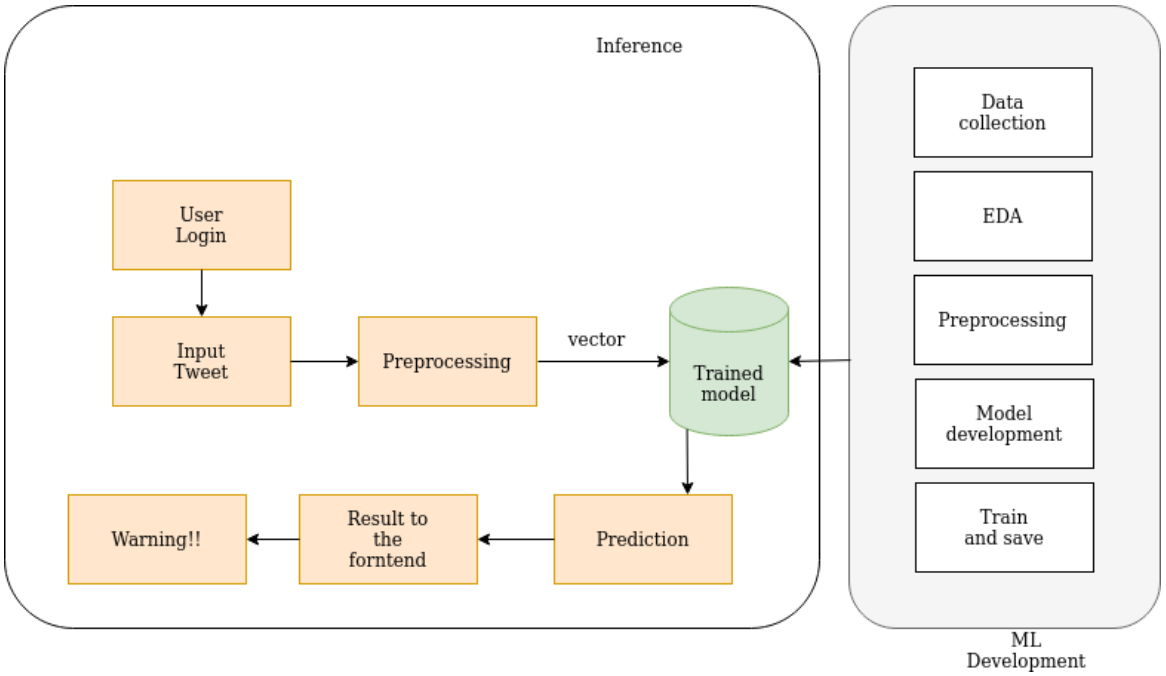


Figure 5.1 : System Architecture.

The Machine Learning development part has the processes right from the data collection to the model training and saving. This trained model after proper testing and evaluation will bedeployed at the backend of the web app or teh twitter clone developed in HTML. This part is called the inference part.

Data Cleaning:

The data collection has the below steps in it.

1. Convert to lower case.

   All the tweets in the dataset are converted to lowercase.

2. Remove special characters.

   Special characters in the tweet text string are removed

3. Remove short forms

   All the short form like “don’t” are replaced bythe corresponding full form (do not).

4. Remove stop words

Stop words are common words like ‘the’, ‘and’, ‘I’, etc. that are very frequent in text, and so don’t convey insights into the specific topic of a document. We can remove these stop words from the text in a given corpus to clean up the data, and identify words that are more rare and potentially more relevant to what we’re interested in.

1. Remove links
   Remove links and URLs in tweets if any.
2. Remove accent
   Remove english accents from the tweet texts using unidecode.
3. Remove long spaces.
   Remove long spaces in the text and replace them with single space.

Other preprocessing steps:

Label encoding

Label Encoding is a technique that is used to convert categorical columns into numerical onesso that they can be fitted by machine learning models which only take numerical data. It is an important preprocessing step in a machine-learning project. Label encoding converts the categorical data into numerical ones, but it assigns a unique number(starting from 0) to each class of data. This may lead to the generation of priority issues during model training of data sets. A label with a high value may be considered to have high priority than a label having a lower value.

Exploratory data analysis (EDA):

EDA helps to visualize the use and distribution of words among each category of cyberbullying in the dataset. EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

TFIDF Vectorizer

TF-IDF stands for Term Frequency Inverse Document Frequency of records. It can be defined as the calculation of how relevant a word in a series or corpus is to a text. The meaning increases proportionally to the number of times in the text a word appears but is compensated by the word frequency in the corpus (data-set).

PCA (Principal Component Analysis)

Principal component analysis, or PCA, is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data points much easier and faster for machine learning algorithms without extraneous variables to process. So, to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

Steps in PCA

Step 1: Standardization

The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis. More specifically, the reason why it is critical to perform standardization prior to PCA, is that the latter is quite sensitive regarding the variances of the initial variables. That is, if there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with small ranges (for example, a variable that ranges between 0 and 100 will dominate over a variable that ranges between 0 and 1), which will lead to biased results. So, transforming the data to comparable scales can prevent this problem. Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

Step 2: Covariance Matrix computation

The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them. Because sometimes, variables are highly correlated in such a way that they contain redundant information. So, in order to identify these correlations, we compute the

covariance matrix. The covariance matrix is a p × p symmetric matrix (where p is the number of dimensions) that has as entries the covariances associated with all possible pairs ofthe initial variables.

Step 3: Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components

Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the principal components of the data. Before getting to the explanation of these concepts, let's first understand what do we mean by principal components. Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components.

Step 4: Feature Vector

As we saw in the previous step, computing the eigenvectors and ordering them by their eigenvalues in descending order, allow us to find the principal components in order of significance. In this step, what we do is, to choose whether to keep all these components or discard those of lesser significance (of low eigenvalues), and form with the remaining ones a matrix of vectors that we call Feature vector. So, the feature vector is simply a matrix that has as columns the eigenvectors of the components that we decide to keep. This makes it the first step towards dimensionality reduction, because if we choose to keep only p eigenvectors (components) out of n, the final data set will have only p dimensions.

Step 5: Recast the Data Along the Principal Components Axes

In the previous steps, apart from standardization, you do not make any changes on the data, you just select the principal components and form the feature vector, but the input data set remains always in terms of the original axes (i.e, in terms of the initial variables). In this step, which is the last one, the aim is to use the feature vector formed using the eigenvectors of the covariance matrix, to reorient the data from the original axes to the ones represented by the principal components (hence the name Principal Components Analysis). This can be doneby multiplying the transpose of the original data set by the transpose of the feature vector.

The Machine Learning models

Decision Tree Classifier

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

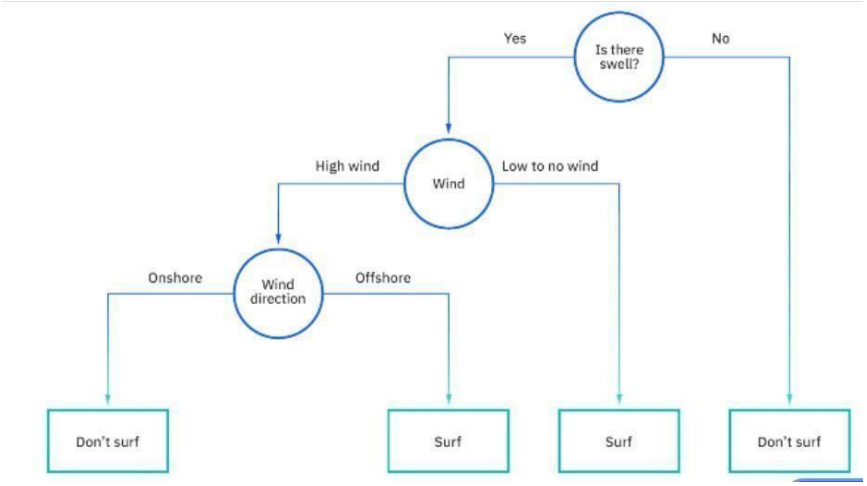


Figure 5.2: Decision Tree

As you can see from the diagram above, a decision tree starts with a root node, which does not have any incoming branches. The outgoing branches from the root node then feed into the internal nodes, also known as decision nodes. Based on the available features, both node types conduct evaluations to form homogenous subsets, which are denoted by leaf nodes, or terminal nodes. The leaf nodes represent all the possible outcomes within the dataset. As an example, let's imagine that you were trying to assess whether or not you should go surf, you may use the following decision rules to make a choice:

Types of Decision Trees Types of decision trees are based on the type of target variable wehave. It can be of two types:

1. Categorical Variable Decision Tree: Decision Tree which has a categorical target variable then it is called a Categorical variable decision tree.
2. Continuous Variable Decision Tree: Decision Tree has a continuous target variable then it is called Continuous Variable Decision Tree.

Steps in the algorithm:

- It begins with the original set S as the root node.

- On each iteration of the algorithm, it iterates through the very unused attribute of the set S and calculates Entropy(H) and Information gain(IG) of this attribute. It then selects the attribute which has the smallest Entropy or Largest Information gain.
- The set S is then split by the selected attribute to produce a subset of the data. The algorithm continues to recur on each subset, considering only attributes never selected before.

Random Forest Classifier

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the predictionfrom each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem ofoverfitting.

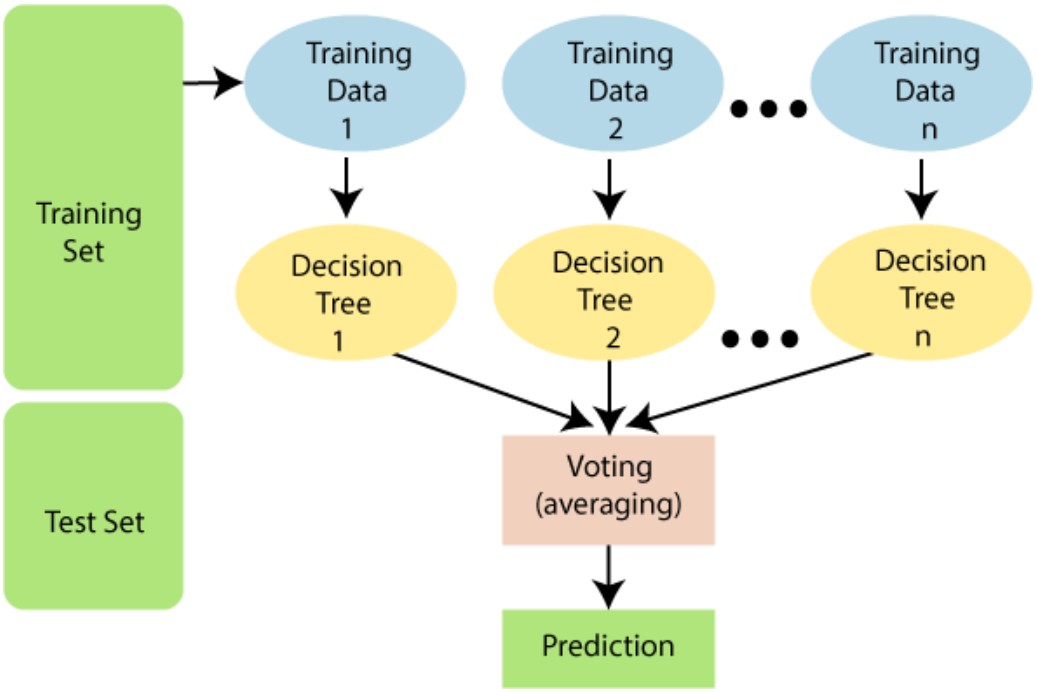


Figure 5.3: Random forest

Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for abetter Random forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Why use Random Forest?

Below are some points that explain why we should use the Random Forest algorithm:

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

How does Random Forest algorithm work?

Random Forest works in two-phase first is to create the random forest by combining Ndecision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the newdata points to the category that wins the majority votes.

Advantages of Random Forest

- Random Forest is capable of performing both Classification and Regression tasks.

- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

Disadvantages of Random Forest

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

XGBoost Algorithm

Ever since its introduction in 2014, XGBoost has been lauded as the holy grail of machine learning hackathons and competitions. From predicting ad click-through rates to classifying high energy physics events, XGBoost has proved its mettle in terms of performance – and speed. The beauty of this powerful algorithm lies in its scalability, which drives fast learning through parallel and distributed computing and offers efficient memory usage.

XGBoost is an ensemble learning method. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models. The models that form the ensemble, also known as base learners, could be either from the same learning algorithm or different learning algorithms. Bagging and boosting are two widely used ensemble learners. Though these two techniques can be used with several statistical models, the most predominant usage has been with decision trees.

Unique features of XGBoost

- Regularization: XGBoost has an option to penalize complex models through both L1 and L2 regularization. Regularization helps in preventing overfitting
- Handling sparse data: Missing values or data processing steps like one-hot encoding make data sparse. XGBoost incorporates a sparsity-aware split finding algorithm to handle different types of sparsity patterns in the data
- Weighted quantile sketch: Most existing tree based algorithms can find the split points when the data points are of equal weights (using quantile sketch algorithm). However, they are not equipped to handle weighted data. XGBoost has a distributed weighted quantile sketch algorithm to effectively handle weighted data

- Block structure for parallel learning: For faster computing, XGBoost can make use of multiple cores on the CPU. This is possible because of a block structure in its system design. Data is sorted and stored in in-memory units called blocks. Unlike other algorithms, this enables the data layout to be reused by subsequent iterations, instead of computing it again. This feature also serves useful for steps like split finding and column sub-sampling
- Cache awareness: In XGBoost, non-contiguous memory access is required to get the gradient statistics by row index. Hence, XGBoost has been designed to make optimal use of hardware. This is done by allocating internal buffers in each thread, where the gradient statistics can be stored
- Out-of-core computing: This feature optimizes the available disk space and maximizes its usage when handling huge datasets that do not fit into memory

Front-end development

The front-end has been developed as a twitter clone web app that can be locally hosted. The three pages are developed in HTML and CSS for styling to make it look exactly like Twitter. The trained Gradient Boost model is deployed in the backend and the API communication is set up in python django.

HTML modules:
Index.html - The twitter home page.
Home.html - The landing page.
Landing.html - The login page.

Chapter 6

SYSTEM REQUIREMENTS

The system requirement is not that much for this project as the training has been carried outin Google Colab free version. So no specific hardware is required.

Hardware requirement:

Basic system with intel i3 or above processor.

Software requirement:

IDE used for ML development and training - Google Colab.

Language used for ML development and training - Python 3.7

Front-end development:

Language used - HTML Language

used for styling - CSS

Language and framework used for API connection - Python - Falsk

In addition to this various python libraries like Tensorflow for deep learning are also used.

Chapter 7
MODULE DESCRIPTION

The whole project can be divided into 3 modules. The first one is the data collection and preprocessing module followed by the Machine Learning module and the final application module.

The Data Collection and Preprocessing Module.

The data collection module involves the below tasks.

Data Collection: The data collection process is pivotal for acquiring a diverse and representative dataset encompassing tweets from various users and topics. Utilizing the Twitter API, a large volume of tweets is gathered, ensuring a wide spectrum of content to facilitate effective model training. The collected data is stored in a structured format, enabling seamless access and analysis throughout the project lifecycle.

Basic Statistics Analysis: A fundamental aspect of data exploration involves conducting basic statistical analyses to gain insights into the characteristics and distribution of the dataset. Descriptive statistics such as mean, median, standard deviation, and frequency distributions are computed to understand the central tendency and variability of different tweet attributes. This preliminary analysis sets the foundation for subsequent exploratory data analysis (EDA) and informs the preprocessing strategies.

Exploratory Data Analysis (EDA): EDA delves deeper into the dataset, uncovering patterns, trends, and anomalies that can influence model performance. Visualizations such as histograms, box plots, and word clouds are employed to visualize tweet distributions, sentiment distributions, and word frequencies. EDA also involves investigating relationships between variables, identifying correlations, and discerning potential features relevant to cyberbullying detection.

Data Cleaning: Data cleaning is imperative to ensure the quality and integrity of the dataset. This involves handling missing values, removing duplicates, correcting inconsistencies, andaddressing outliers. Text preprocessing techniques such as tokenization, lowercasing, removal of stop words, punctuation, and special characters are applied to standardize the textual data and prepare it for feature extraction.

Feature Extraction: Feature extraction involves transforming raw tweet data into a structured format suitable for model training. Features such as word frequencies, n-grams, sentiment scores, and syntactic features are extracted to capture relevant information indicative of cyberbullying

behavior. Additionally, domain-specific features such as user metadata and tweet context may be incorporated to enhance model performance.

Preprocessing: Preprocessing encompasses a series of steps to prepare the dataset for input into Machine Learning models. This includes encoding categorical variables, scaling numerical features, and splitting the data into training and testing sets. Furthermore, techniques such as text normalization, vectorization, and dimensionality reduction may be applied to streamline the data and optimize model training efficiency.

Overall, the data collection and preprocessing module lays the groundwork for subsequent model development by ensuring the dataset is comprehensive, clean, and properly formatted for effective Machine Learning-based cyberbullying detection within the Twitter clone application.

The Machine Learning Module.

The Machine Learning module constitutes a pivotal phase in the development of the cyberbullying detection system within the Twitter clone application. This module focuses on the implementation and evaluation of machine learning algorithms, including Random Forest, Decision Tree, and XGBoost, leveraging the scikit-learn library in Python and the Google Colab IDE for seamless development and evaluation.

Algorithm Development: The initial step involves the development of machine learning algorithms tailored to the task of cyberbullying detection. Random Forest, Decision Tree, and XGBoost algorithms are selected for their effectiveness in handling structured data and ability to capture complex patterns inherent in tweet content. These algorithms are instantiated and configured using the scikit-learn library, with parameters tuned to optimize performance.

Training and Evaluation: Following algorithm development, the next phase entails training the models on the preprocessed tweet dataset and evaluating their performance using various metrics. The training process involves feeding the algorithm with labeled tweet data, wherein the model learns to discern between cyberbullying and non-cyberbullying instances. The scikit-learn library facilitates efficient model training with its user-friendly interface and extensive documentation.

Evaluation Metrics: The efficacy of the machine learning algorithms is assessed based on a set of predefined evaluation metrics, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). Accuracy measures the proportion of correctly classified instances, while precision quantifies the ratio of true positive predictions to the total number of positive

predictions. Recall, also known as sensitivity, evaluates the model's ability to correctly identify positive instances from the entire dataset. F1-score represents the harmonic mean of precision and recall, providing a balanced assessment of the model's performance.

Model Optimization: To enhance model performance, hyperparameter tuning techniques such as grid search and randomized search may be employed to find the optimal combination of model parameters. By systematically exploring the hyperparameter space, the models can be fine-tuned to achieve improved predictive accuracy and generalization capabilities. The iterative nature of model optimization ensures that the algorithms are robust and resilient to overfitting or underfitting tendencies.

The Application Module.

The final module, known as the application module, represents the culmination of efforts in the development of the cyberbullying detection system within the Twitter clone application. This phase encompasses the design, development, and deployment of the Twitter clone interface in HTML, integration with the backend Flask framework, and real-time testing of the entire system using the ngrok platform.

Design and Development: The design phase entails creating a user-friendly and intuitive interface that closely resembles the layout and functionality of the original Twitter platform. Using HTML, developers craft web pages comprising elements such as timelines, user profiles, tweet composition boxes, and notification panels. CSS styling is applied to enhance the visual appeal and ensure consistency across different browsers and devices. The design process involves careful consideration of user experience principles to optimize navigation and interaction flows.

In parallel, the backend logic is implemented using the Flask framework in Python. Flask provides a lightweight yet powerful framework for building web applications, offering features such as routing, request handling, and session management. Developers define routes corresponding to different functionalities of the Twitter clone, such as posting tweets, following users, and detecting cyberbullying. Integration with the machine learning module enables real- time analysis of incoming tweets for potential cyberbullying content.

Deployment with ngrok: Once the development phase is complete, the Twitter clone applicationis deployed to a publicly accessible server using the ngrok platform. Ngrok facilitates secure tunneling of web traffic to a local development environment, allowing for external access to the

application without the need for complex network configurations. Developers generate a unique ngrok URL that serves as the public endpoint for accessing the Twitter clone application.

Real-Time Testing: With the application deployed and accessible via the ngrok URL, real-time testing is conducted to validate the functionality and performance of the cyberbullying detection system. Test scenarios involve simulating user interactions, such as posting tweets containing both benign and potentially harmful content. The machine learning module analyzes the incoming tweets in real-time, identifying instances of cyberbullying based on the trained models' predictions.

Chapter 8
IMPLEMENTATION

Data collection and preprocessing

The dataset of tweets has been collected from kaggle.com. It contains 47692 tweet samples. The whole dataset has 6 types of tweets in it as shown below.

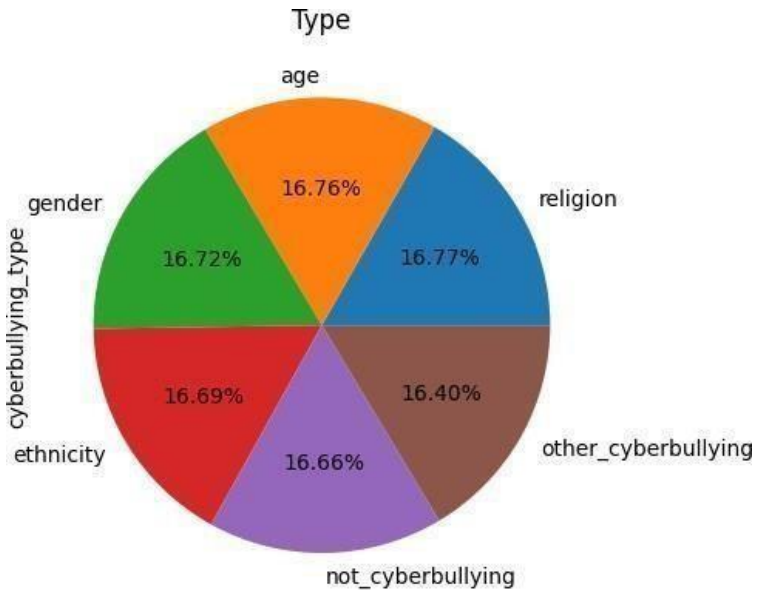


Figure 8.1: Types of tweets in the dataset.

The pie chart of tweet type distribution shows that the classes are equally distributed, so itwill help the ML model we are training learn well.

Number of tweets in each category:

- Religion 7998
- Age 7992
- Gender 7973
- Ethnicity 7961
- Not_cyberbullying 7945
- Other_cyberbullying 7823

Exploratory data analysis (EDA):

The below are the result of the EDA and word cloud generation to know most frequent wordsin each category.

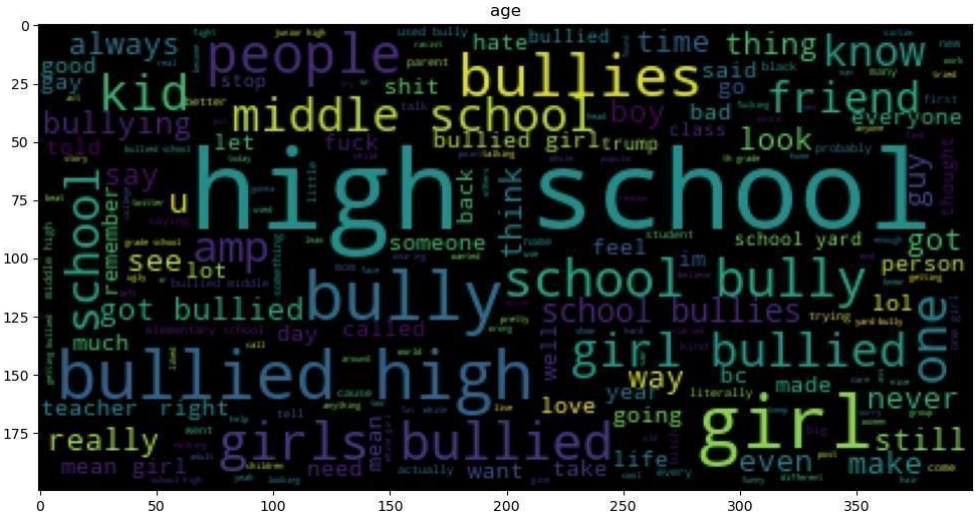


Figure 8.2 : word distribution in age category.

Figure 8.3 : word distribution in ethnicity category.

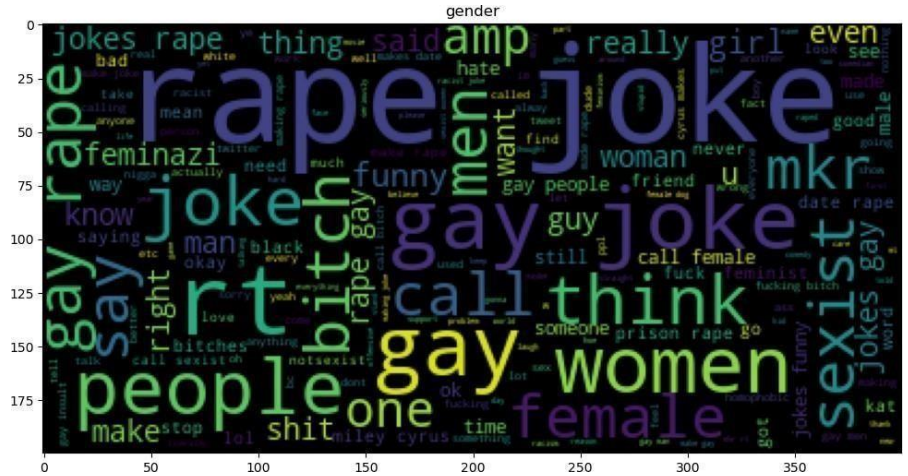


Figure 8.4: word distribution in gender category.

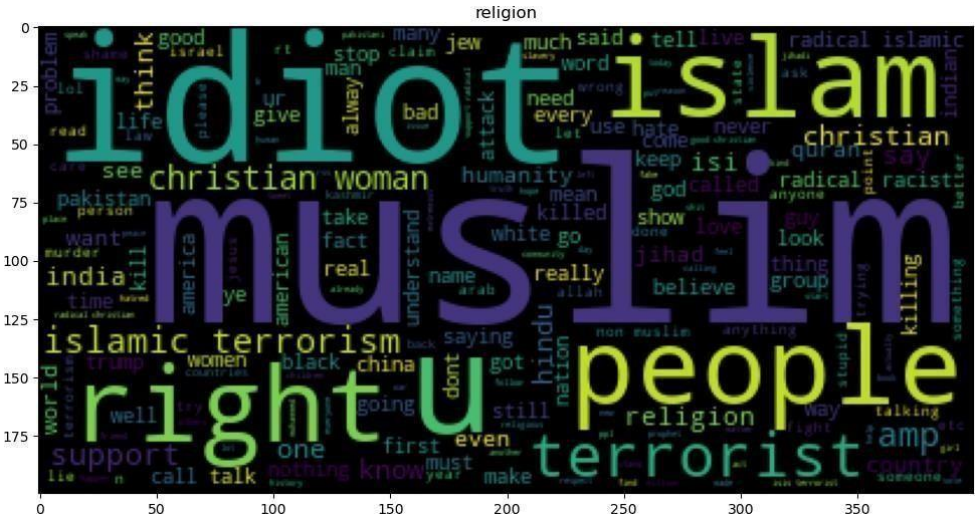


Figure 8.5: word distribution in age category.

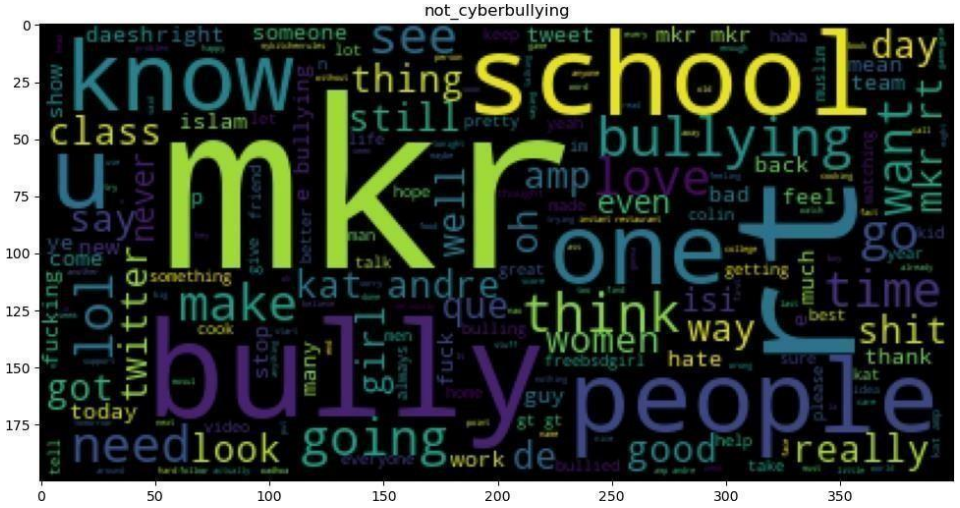


Figure 8.6: word distribution in non bullying category.

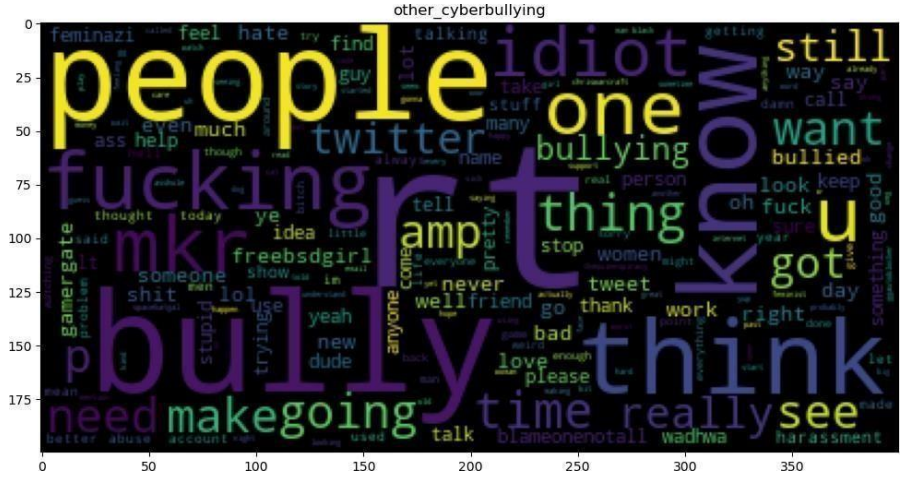


Figure 8.7: word distribution in other bullying categories.

The preprocessed data is split into a train set and test set at ratio 80:20.

The Machine Learning models' results

Decision Tree Classifier

Training result:

- Max depth = 100
- Training data = 38153
- Testing data = 9539 @ 80:20
- Training accuracy achieved = 96.8%
- Testing accuracy achieved = 71.1%

Random Forest Classifier

Training result:

- Number of trees = 100
- Training data = 38153
- Testing data = 9539 @ 80:20
- Training accuracy achieved = 96.80%
- Testing accuracy achieved = 78.0%

XGBoost Algorithm

Training result:

- Number of trees = 125
- Training data = 38153
- Testing data = 9539 @ 80:20
- Training accuracy achieved = 94.2%
- Testing accuracy achieved = 78.9%

Offline Testing.

The final model; teh XGBoost model has been tested in the Google Colab itself with random tweet texts. And the samples of some of the results are given below.

Figure 8.8: Test result 1

Figure 8.9: Test result 2

Frontend Design.

The three page frontend is designed in HTML and CSS.

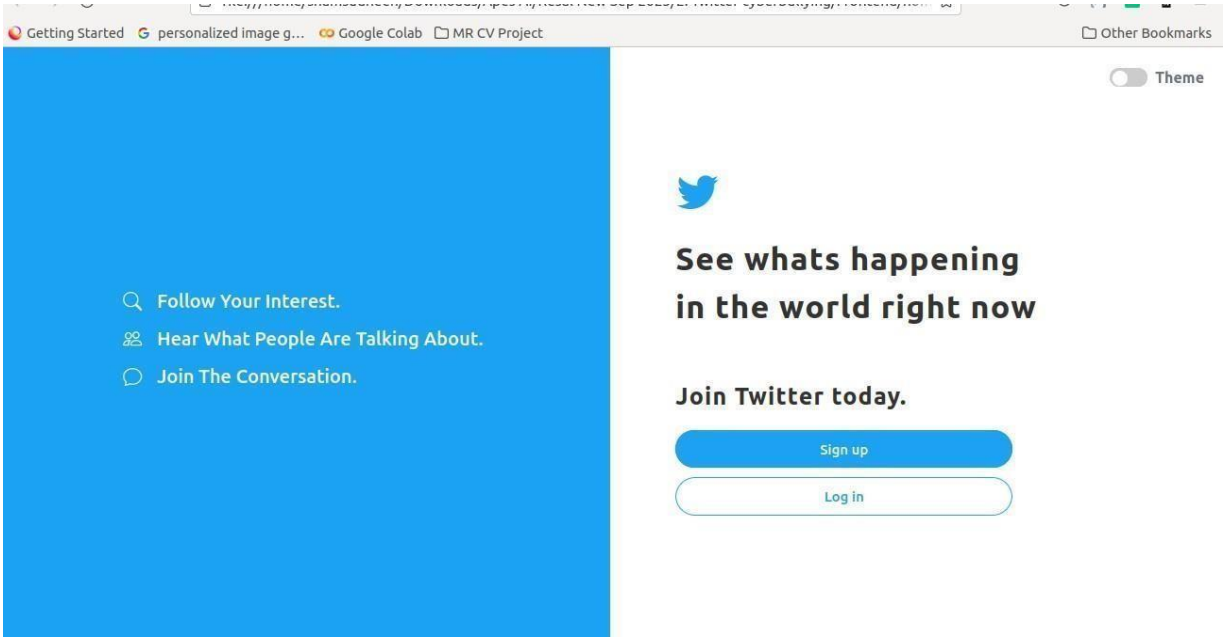


Figure 8.10: The landing page

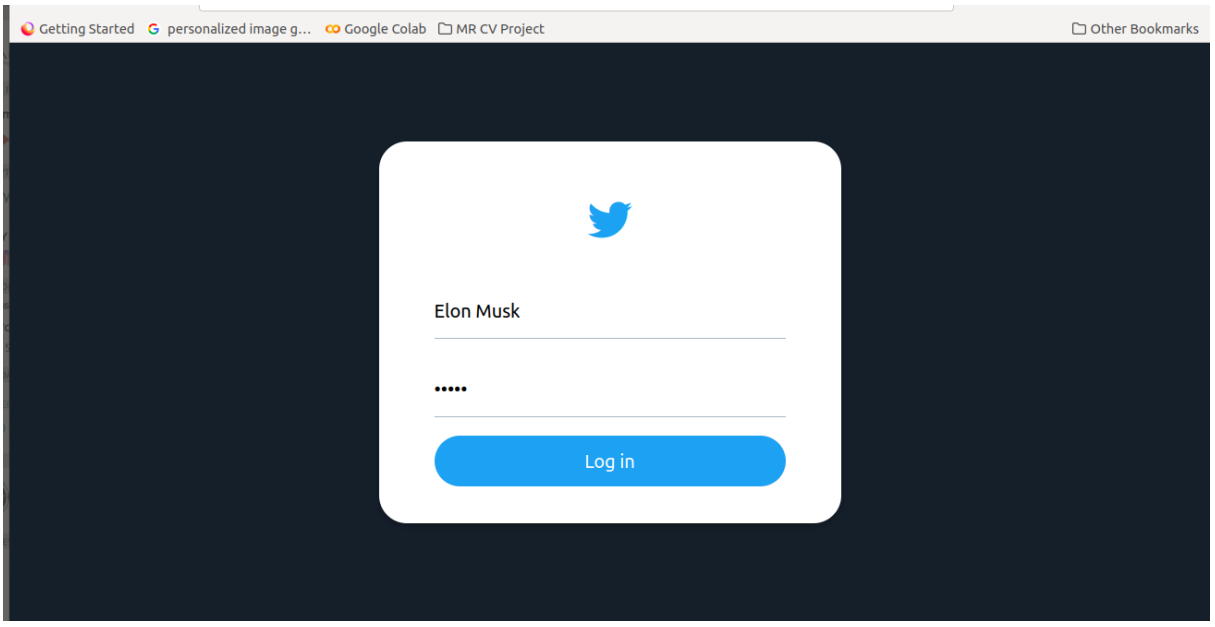


Figure 8.11: The login page

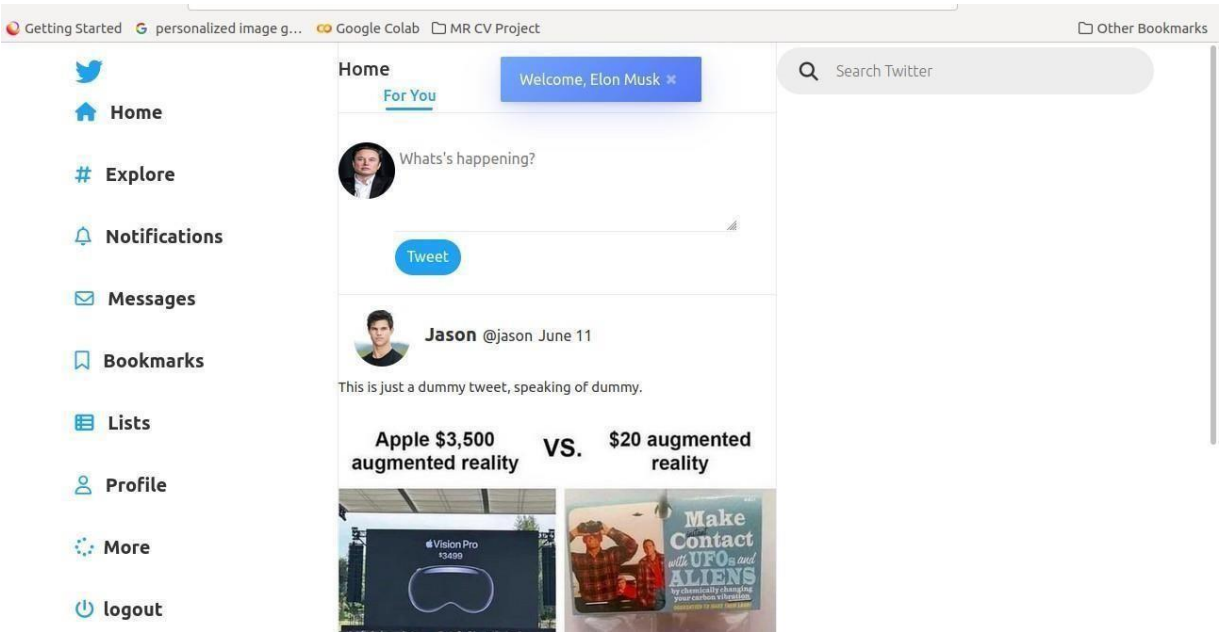
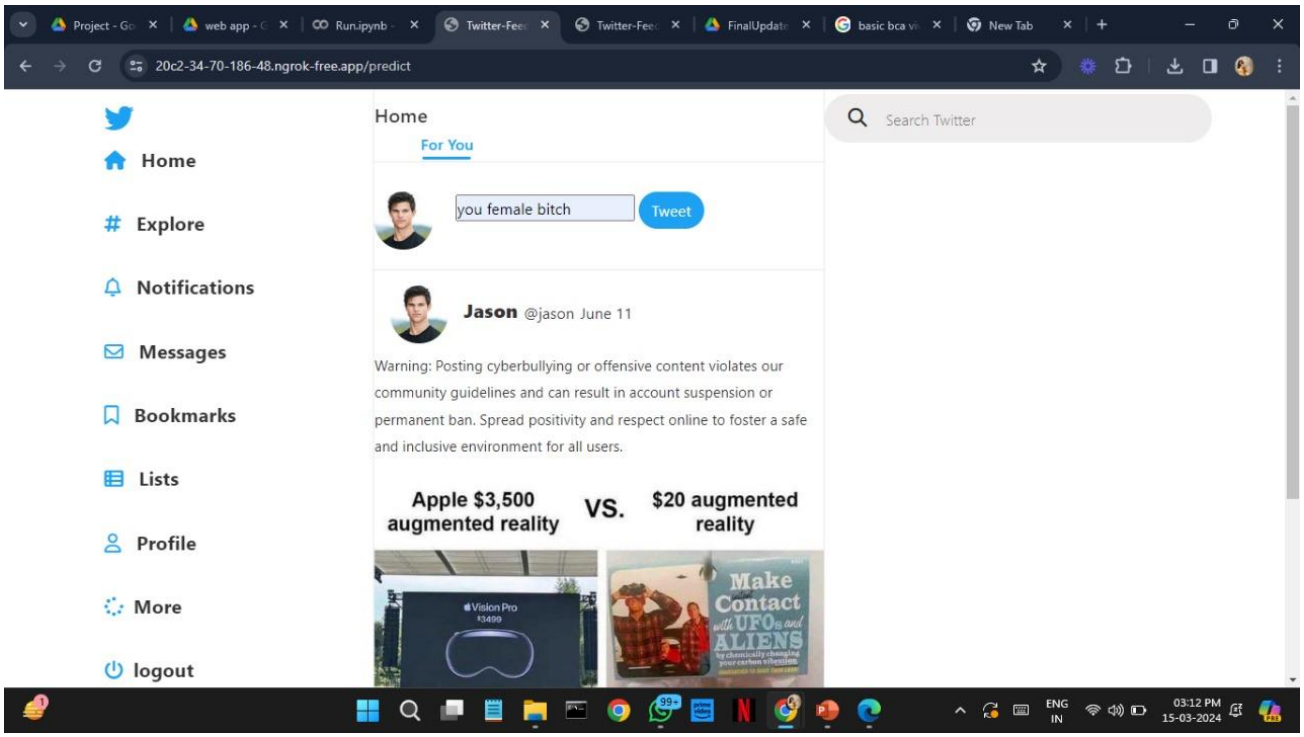
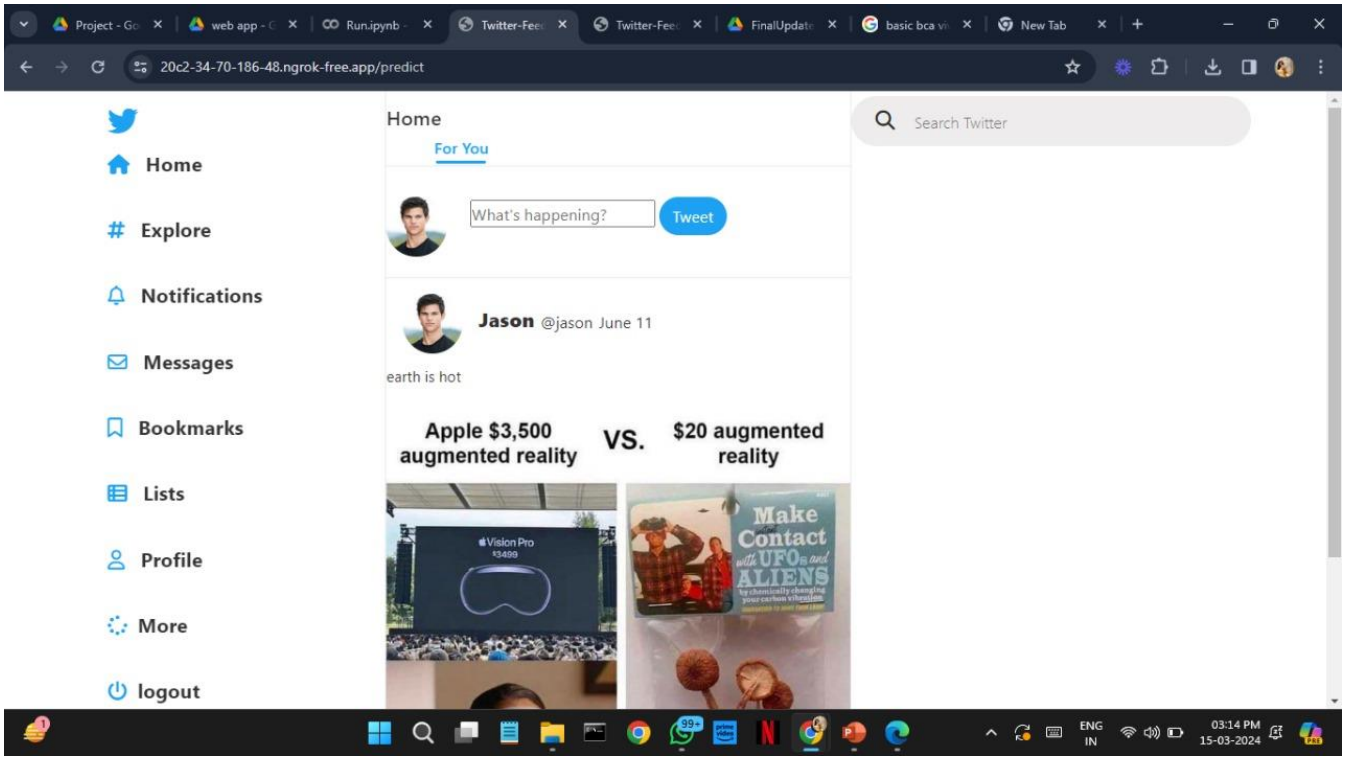


Figure 8.12: The main page

The final test result.



8.13 The Final Test Result



8.14 The Final Test Result 2

Chapter 9

CONCLUSION

Cyberbullying is a serious social issue that has increased since the inception and popularization of social media. It has a number of emotional, mental, and physical effects on victims including children. Social media platforms including Twitter are working everyday to fight cyberbullying by developing advanced and efficient technologies. A twitter cyberbullying tweet data has been collected in this project. The collected data passed through a set of Natural Language Processing (NLP)-based operations to extract features from the tweets Machine learning models are trained. The XGBoost model achieved better testing accuracy in spotting cyberbullying tweets. A Twitter clone web app has been developed to test the model trained and test successfully with real-time alert. The machine learning models benefited because of adding the NLP-based preprocessing operations. The accuracy can be increased by adding more data and adding more preprocessing operations.

REFERENCES

1. R. R. Dalvi, S. Baliram Chavan and A. Halbe, "Detecting A Twitter Cyberbullying Using Machine Learning,"
2. T. Balet, Q. Vo, O. Salem and A. Mehaoua, "Cyberbullying Detection on tweets from Twitter using Machine Learning Algorithms," 2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS), Valencia, Spain, 2023, pp. 177-182, doi: 10.1109/ICCNS58795.2023.10193450.
3. S. A. Mathur, S. Isarka, B. Dharmasivam and J. C. D., "Analysis of Tweets for Cyberbullying Detection,"
4. Salawu, S., Lumsden, J., & He, Y. (2021). A Large-Scale English Multi-Label Twitter Dataset for Cyberbullying and Online Abuse Detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (pp. 146–156). Association for Computational Linguistics.
5. Muneer A, Alwadain A, Ragab MG, Alqushaibi A. Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT. Information. 2023; 14(8):467. https://doi.org/10.3390/info14080467
6. Fati SM, Muneer A, Alwadain A, Balogun AO. Cyberbullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continuous Bag of Words Feature Extraction. Mathematics. 2023; 11(16):3567. https://doi.org/10.3390/math11163567
7. Parikh, R., Dalvi, A. (2024). Identifying Instances of Cyberbullying on Twitter Using Deep Learning. In: Kulkarni, A.J., Cheikhrouhou, N. (eds) Intelligent Systems for Smart Cities. ICISA 2023. Springer, Singapore. https://doi.org/10.1007/978-981-99-6984-5_6
8. Raj M, Singh S, Solanki K, Selvanambi R. An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques. SN Comput Sci. 2022;3(5):401. doi: 10.1007/s42979-022-01308-5. Epub 2022 Jul 26. PMID: 35911437; PMCID: PMC9321314.
9. B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif and H. D. E. Al-Ariki, "DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform," in IEEE Access, vol. 10, pp. 25857-25871, 2022, doi: 10.1109/ACCESS.2022.3153675.

10. Hasan MT, Hossain MAE, Mukta MSH, Akter A, Ahmed M, Islam S. A Review on Deep-Learning-Based Cyberbullying Detection. Future Internet. 2023; 15(5):179. https://doi.org/10.3390/fi15050179
11. Obamiyi, S. E., Badeji-Ajisafe, B., Oguntimilehin, A., Adefehinti, T., Abiola, O., & Okebule, T. (2023). An Ensemble Approach to Cyberbullying Detection and Prevention on Social Media. ABUAD International Journal of Natural and Applied Sciences, 3(2), 47-52. https://doi.org/10.53982/aijnas.2023.0302.07-j
12. Nitya Harshitha, T., Prabu, M., Suganya, E., Sountharrajan, S., Bavirisetti, D., Gadde, N., & Uppu, L. (2024). ProTect: a hybrid deep learning model for proactive detection of cyberbullying on social media. Frontiers in Artificial Intelligence, 7.
13. Selvakumar, M. ., & Kathiravan, A. V. . (2023). Deep Learning based Densenet Convolution Neural Network for Community Detection in Online Social Networks. International Journal on Recent and Innovation Trends in Computing and Communication, 11(8s), 202–214. https://doi.org/10.17762/ijritcc.v11i8s.7191

APPENDICES

Source Code

```
<html lang="en">
  <head>
    <meta charset="utf-8" />
    <meta name="viewport" content="width=device-width, initial-scale=1" />
    <meta name="theme-color" content="#000000" />
    <meta name="description" content="Twitter Clone" />
    <title>Twitter-Feed</title>
    <link rel="icon" href="../images/twitter.png" />

    <link
      rel="stylesheet"
      href="https://cdnjs.cloudflare.com/ajax/libs/jquery-modal/0.9.2/jquery.modal.min.css"
      integrity="sha512-
      T3VL1q6jMUIzGLRB9z86oJg9PgF7A55eC2XkB93zyWSqQw3
      Ju+6IEJZYBfT7E9w OHM7HCMCOZSpcssxnUn6AeQ=="
      crossorigin="anonymous"
      referrerpolicy="no-referrer"
    />
   <link href="../static/css/bootstrap.min.css" rel="stylesheet" />
    <link
      rel="stylesheet"
      href="https://cdnjs.cloudflare.com/ajax/libs/font-awesome/6.4.0/css/all.min.css"
      integrity="sha512-ie
      cdLmaskl7CVkqkXNQ/ZH/XLlvWZOJyj7Yy7tc
      enmpD1ypASozpmT/E0iPtmFIB46ZmdtAc9eNBvH0H/ZpiBw=="
      crossorigin="anonymous"
      referrerpolicy="no-referrer"
      />
```

```
<p class="option">
 <a><i class="fa-solid fa-spinner"></i>More</a>
</p>
<p class="option">
 <a href="home.html" id="logoutBtn"
  ><i class="fa-solid fa-power-off"></i>logout</a
  >
 </p>
 <button class="btn btn-lg btn-primary">Tweet</button>
 </div>
</div>
<div class="col-5 feed">
<div class="header">
<h5>Home</h5>
<div class="d-flex ms-5">
 <h6>For You</h6>
 </div>
 </div>
<div class="post">
<div class="avatar">
<img src="../static/jason.jpg" />
 </div>
 <div class="textbox">

 <form action="{{ url_for('predict')}}"method="post">
 <input type="text" name="enginesize" placeholder="What's happening?"
  required="required" />

 <button class="btn">Tweet</button></a
  >
 </form>
 <div id="ex1" class="modal">
  <div>

  <img
```

```
src="../static/warning.png"
class="m-auto d-block mt-2 mb-3"
width="50"
 alt="warning"
  />
 </div>
 <p class="text-danger fw-bold" style="text-align: justify">
   The post you are trying to share cannot be posted because it
   violates Twitter's cyberbullying policies. We prioritize
   maintaining a safe and respectful environment for all users.
   If you have any questions or concerns, please review our
   community guidelines or contact our support team for further
   assistance. Thank you for your understanding.
   </p>
   <!-- <a href="#" rel="modal:close">Close</a> -->
   </div>
   </div>
   </div>

  <div class="tweets">
  <div
  class="profile-pic d-flex justify-content-start align-items-center"
   >
 <img src="../static/jason.jpg" />
 <div class="tweet-heads ms-3 d-flex">
 <h4>Jason</h4>
 <span>@jason</span>
 <span>June 11</span>
 </div>
 </div>
 <div class="tweet-content">
 <p> {{prediction_text}}</p>
 <img src="../static/post.jpg" alt="tweet" class="img-fluid" />
 </div>
```

```
 <div class="tweet-icons">
 <i class="fa fa-comment" aria-hidden="true"></i>
 <i class="fa fa-heart" aria-hidden="true"></i>
 <i class="fa fa-retweet" aria-hidden="true"></i>
</div>
</div>
</div>
<div class="col-4 right-bar">
<div class="search-box">
<i class="fa fa-search" aria-hidden="true"></i>
<input type="text" placeholder="Search Twitter" />
</div>
</div>
</div>
</div>
</section>
<script src="https://ajax.googleapis.com/ajax/libs/jquery/3.5.1/jquery.min.js"></script>
<script
src="https://cdnjs.cloudflare.com/ajax/libs/jquery-modal/0.9.2/jquery.modal.min.js"
integrity="sha512-
ztxZscxb55lKL+xmWGZEbBHekIzy+1qYKHGZTWZYH1GUwxy0hiA18l
W6ORIMj4DHRgvmP/qGcvqwEyFFV7OYVQ=="
crossorigin="anonymous"
referrerpolicy="no-referrer"
></script>
<script src="script/bootstrap.bundle.min.js"></script>
<script src="https://cdn.jsdelivr.net/npm/toastify-js"></script>
<script>

</script>
</body>
</html>
```

```
  # Load the XGBoost model
  model = pickle.load(open("XGBoost_model.pickle", "rb"))

# The final function to make predictions
def listToString(s):

# initialize an empty string
str1 = ""

# traverse in the string
for ele in s:
str1 += ele

# return string
return str1

def detect(tweet):
 inputs=[clean_text(tweet)]    # Apply the cleaning function
 inputs=tfidf.transform(inputs).toarray()   # Apply TFIDF
 inputs=pca.transform(inputs)     # Apply PCA
 a=model.predict(inputs)[0]    # Ask model to predict
 pred=lenc.classes_[a]  # Apply label encoder

 if a==0 or a==1 or a==2 or a==4 or a==5:  # If any of the cyberbullying types
 print('Cyber bullying')
 elif a==3:                  # If not cyberbullying
 print('Not cyber  bullying')
```