A

PROJECT REPORT

ON

CYBERBULLYING DETECTION AND PREVENTION IN WEB CHAT APPLICATION USING SUPPORT VECTOR MACHINE

Submitted by

B190138518:- Siddhi Damale

B190138528:- Apurva Hyalij

B190138529:- Rohini Jadhav

B190138536:- Pooja Khalkar

Guided by

Prof. Yogita H. Khairnar

In partial fulfillment for the award of a degree

Of

Bachelor of Engineering

Of

Savitribai Phule Pune University

IN

DEPARTMENT OF

INFORMATION TECHNOLOGY

K.K. WAGH INSTITUTE OF ENGINEERING EDUCATION & RESEARCH, NASHIK-3

2022-23

A

PROJECT REPORT

ON

CYBERBULLYING DETECTION AND PREVENTION IN WEB CHAT APPLICATION USING SUPPORT VECTOR MACHINE

Submitted by

B190138518:- Siddhi Damale

B190138528:- Apurva Hyalij

B190138529:- Rohini Jadhav

B190138536:- Pooja Khalkar

Guided by

Prof. Yogita H. Khairnar

In partial fulfillment for the award of a degree

Of

Bachelor of Engineering

Of

Savitribai Phule Pune University

IN

DEPARTMENT OF

INFORMATION TECHNOLOGY

K.K. WAGH INSTITUTE OF ENGINEERING EDUCATION & RESEARCH, NASHIK-3

2022-23

DEPARTMENT OF INFORMATION TECHNOLOGY

CERTIFICATE

This is to certify that the project report entitled

CYBERBULLYING DETECTION AND PREVENTION IN WEB CHAT APPLICATION USING SUPPORT VECTOR MACHINE

Submitted by

B190138518:- Siddhi Damale

B190138528:- Apurva Hyalij

B190138529:- Rohini Jadhav

B190138536:- Pooja Khalkar

is a bonafide work carried out by them under the supervision of Prof. Yogita H. Khairnar and it is approved for the partial fulfillment of the requirement of Savitribai Phule Pune University for the award of the Degree of Bachelor of Engineering (Information Technology).

This project report has not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. Yogita H. Khairnar	**Prof. Dr. Preeti D. Bhamre**
Internal Guide	Head of Department
Department of Information Technology	Department of Information Technology

External Examiner

Date:

Place:

Date:

ACKNOWLEDGEMENT

Motivation and guidance is the key to success. We would like to thank all the sources of motivation and guidance with due respect and gratitude. It gives us great pleasure and satisfaction in presenting this project report on “Cyber Bullying Detection and Prevention in Web Chat Application using Support Vector Machine".

We would like to express our deep sense of gratitude towards the Information Technology Department, **Prof. Dr. Preeti D. Bhamre** and **Prof. Yogita H. Khairnar** to encourage us to go ahead and for her continuous guidance. We also want to thank her for all her assistance and guidance in preparing the report.

Finally We would like to extend our thanks to all teaching and non-teaching staff members of the Information Technology Department for their cooperation in this project.

Thanking you,

Siddhi Ravindra Damale

Apurva Dattatray Hyalij

Rohini Rajaram Jadhav

Pooja Vishnu Khalkar

Abstract

In todays life, Social media is becoming increasingly popular, and its effects are being felt more and more. Cyberbullying has emerged as a critical affliction for children, youngsters, and teenagers.

Cyberbullying refers to intentional actions performed by an individual or a group of people via digital communication methods, such as sending messages and posting comments against a victim. Cyberbullying leads to an increase in stress and anger in a person. Preventing cyberbullying is difficult because it can be challenging to protect victims from these activities. Although several techniques have been developed to detect social media bullying, there is currently no automatic system for detecting it in live chat applications. Therefore, a project has been initiated to identify alarming words or dangerous content in live chat applications. A classifier, a Support Vector Machine is used for training and testing the dataset to predict abusive words in live conversation.

Keywords: Cyber Bullying, Machine Learning, Real-time Web Chat Application, Offensive Words, Support Vector Machine

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

This chapter provides an introduction to the concept of cyberbullying with the scope and objectives of the project.

The arrival of the internet has led to the widespread use of social media platforms like Facebook, Twitter, Instagram, and WhatsApp as the primary means of communication. Messaging has become an essential tool in almost all sectors, including education, business, and socialization. However, it has also created new possibilities for harmful activities. Evidence suggests that messaging can give rise to an issue known as cyberbullying.

Cyberbullying refers to the intentional use of messaging on social networking services to cause emotional, mental, or even physical harm to others. It can lead to psychological problems such as loneliness, low self-esteem, social anxiety, depression, and even suicide. Reports indicate that around 36% of Indian children are affected by cyberbullying. As the occurrence of cyberbullying continues to increase, it is vital to study effective ways to detect and prevent it in real time. Blocking messages alone is insufficient in preventing such incidents. Instead, monitoring, processing, and analyzing text messages in real-time is necessary to make informed decisions and prevent harm to victims.

Due to the aforementioned issues, several studies have been conducted to explore effective techniques for detecting cyberbullying. While manual detection is the most accurate method, it is rarely employed due to the significant time and resource requirements. Therefore, the importance of developing an automatic cyberbullying detection system has increased.

Although considerable research has been conducted on cyberbullying detection systems, it remains a pressing concern, and current approaches still have limitations, especially when dealing with large volumes of data. Various types of social networking services (SNS) can represent different forms or patterns of data. A supervised machine-learning approach can be applied to address cyberbullying from various perspectives. Our primary objective is to develop a classification model that can predict text messages and prevent cyberbullying in real time. The detection process is automated, with the abusive language being quickly identified and a warning message displayed to the user who used the abusive language.

1.1 SCOPE :

1. The System only focuses on the English language.
2. The system can recognize only proper and formal text.

1.2 OBJECTIVES:

1. The system will detect and prevent the cyberbullying event to deteriorate.
2. Web chat application will be used by people for safer and more secure communication with their friends.
3. Cyberbullying leads to an increase in stress and anger in a person, Being a victim of cyberbullying also affected student's grades and hence if peoples use our system they will live a healthy social life.

Chapter 2

LITERATURE REVIEW

This chapter discusses the literature review for the project.

Cyberbullying is an act of threatening, harassing, or bullying someone through modern ways of communicating with each other and with anybody/everybody in the world via social media apps/sites. Cyberbullying is not just limited to creating a fake identity and publishing/posting some embarrassing photo or video, or unpleasant rumors about someone but also giving them threats. The impacts of cyberbullying on social media are horrifying, sometimes leading to the death of some unfortunate victims. The behavior of the victims also changes due to this, which affects their Emotions, self-confidence, and a sense of fear is also seen in such people.

Al-Garadi et al. [1] comprehensively reviewed cyberbullying prediction models and identified the main issues related to the construction of cyberbullying prediction models in social media. The paper provides insights into the overall process for cyberbullying detection and mainly focuses on feature selection algorithms and the use of various machine learning algorithms for the prediction of cyberbullying behavior. The authors found that SVM was an effective and efficient algorithm for developing cyberbullying detection models. The model was trained using data containing cyberbullying extracted from a social network site. The paper reviews four aspects of detecting cyberbullying messages using machine learning approaches: data collection, feature engineering, construction of cyberbullying detection models, and evaluation of constructed cyberbullying detection models.

Shutonu Mitra et al. [2] The research paper proposes a conceptual framework for detecting and preventing cyberbullying on social media. The framework consists of three modules, namely the User Interaction module, the Decision-making module, and the Analysis module. The User Interaction module is responsible for collecting data from users, while the Decision-making module analyses the collected data and makes decisions based on the analysis. Finally, the Analysis module is responsible for processing the data and generating reports. The framework utilizes various machine learning algorithms to analyze and detect cyberbullying behavior on social media. The research paper aims to provide an effective solution for preventing cyberbullying in social media using machine learning algorithms.

In recent years, several studies have been conducted on the analysis, detection,

and prevention of online bullying using text-mining techniques for classifying conversations. One of these studies was conducted by John Hani et al. [3], where they proposed a classification model to detect and prevent social media bullying using neural networks and SVM. They collected their dataset from Kaggle and divided their proposed model into three major steps:

1. Pre-processing Steps:
    - Lowering Text
    - Tokenization
    - Stop words
    - Word correction
2. Feature Extraction:
    - Tokenization
    - Lowering Text
3. Classification:
    - SVM (Support Vector Machine) and NN classifiers were used for classification.

The combination of features with an SVM classifier for predicting the class labels produced the best results. The experiments showed the models ability to detect Insult and Hate content in text. The f1- score of the Insult and Hate class label showed a high score respectively 95.4 [4]. The Support Vector Machine achieved the highest accuracy i.e. 71.25%, while Naive Bayes achieved 52.70% accuracy. The SVM algorithm achieved the highest precision value i.e. 71%, while NB achieved 52% precision. Also, SVM has achieved higher recall and f-score values than Naive Bayes [5].

In recent years, various studies have been conducted to explore the effectiveness of different machine-learning algorithms in detecting cyberbullying using text data from social media platforms. Karthik [6] used multi-classifiers such as Naïve Bayes, JRip, J48, and SMO with YouTube comments. Similarly, Vinita [7] employed LDA to extract features and used the weighted term frequency-inverse document frequency (TF-IDF) function to improve the classification with datasets from Kongregate, Slashdot, and MySpace. Homa [8] utilized the Support Vector Machines classifier with datasets from Instagram.

The proposed project involves building a system using Python and Django frameworks to detect cyberbullying in live conversations. The first step is to search and download the required dataset, which will be pre-processed and used to train the model. Support Vector Machine will be used to generate the model. A web-based application will be created using the Django framework, and the generated model will be applied to live conversations to determine whether the messages are instances of cyberbullying or not.

2.1 EXISTING SYSTEM

This chapter gives an overview of the existing systems that are developed to detect cyberbullying. Figure 2.1 shows the interface of the "ReThink" keyboard, which is the keyboard that detects bullying words and gives an alert to the user who has typed it.

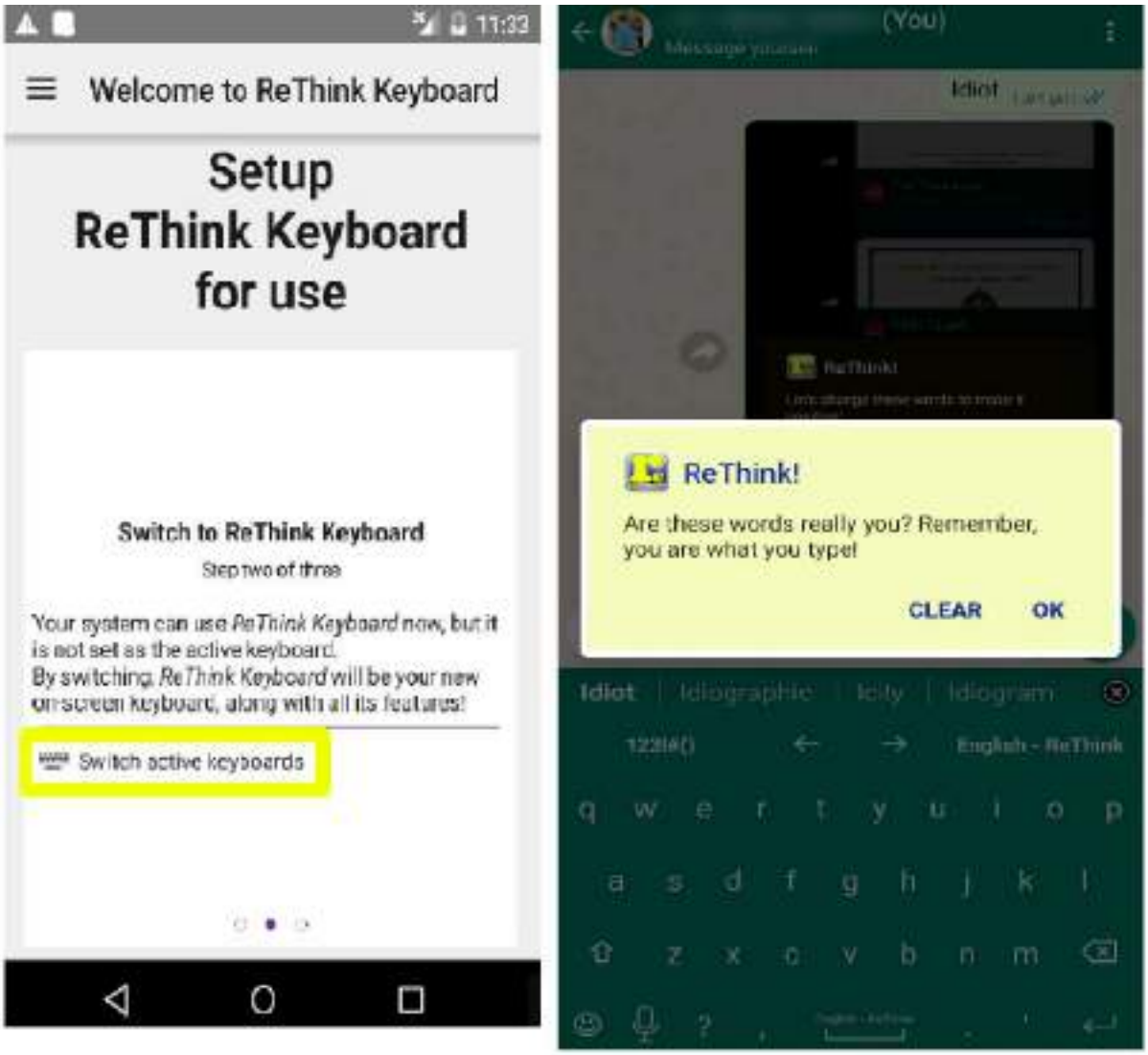

Figure 2.1: Existing System

"ReThink Before The Damage is Done" is an innovative and patented technology that effectively detects if the word that user has entered is cyberbullying or not and stops online hate before the damage is done. This keyboard shows the popup/alert if someone typed offensive words. It suggests clearing that message and rethink about it, but the user also has the option to choose "OK" and continue chatting.

The next topic discusses the work done related to cyberbullying issues.

2.2 RELATED WORK DONE

There are a few applications that are developed to detect cyberbullying, Bully Alert and the Bull Stop application are two of them which are discussed in this chapter.

Bull Stop

Bull Stop application is a mobile application developed for Android devices as part of a Ph.D. study at the University of Aston, United Kingdom (Salawu, 2020). The application is designed to help young people combat cyberbullying proactively, with its target audience being children aged 13 and above. The application works by using a deep learning-based algorithm to analyze messages and flag any offensive content such as cyberbullying, abuse, or insults. For this application, if the application detects some offensive content in the user's social media account, it will automatically delete the messages to prevent them from reaching the user. Bull Stop application also allows users to block other users' social media from the application directly. This application currently only works on Twitter's social media accounts with plans for Facebook and Instagram in the future. This application's advantage is that it is easy to set up and will automatically delete the cyberbullying content just by running this application in the background. Another advantage of this application will also be that this application does not allow parents to monitor their children's social media accounts to help preserve the user's social media account privacy. However, the disadvantage of this application is that it does not show the messages that are deleted, which could delete a message accidentally if the algorithm flagged the messages wrongly. Another disadvantage of this application is that the user won't see the cyberbullying context as the message will be deleted before it reaches the user.

Bully Alert

Bully Alert application is a mobile application developed by CU Cyber Safety Lab as a research project for the University of Colorado Boulder (Lab 2018). Bully Alert application is designed for Android and can be downloaded from Google Play Store. Anyone can use this application to observe another person's social media account to get a notification from their user's profile. This application, as of now, can only be used for an Instagram account. To use this application, the user must input the user's Instagram profile name in the application. After that, the application will start observing the Instagram profile inputted into the application and then send a notification if a cyberbullying case is happening on that Instagram profile.

The advantage of using this application is that it is easy to use and only requires the profile name of the person's Instagram to start observing the profile movement. The disadvantage of this application is that it can only be used if the user knows the user's Instagram profile name and the warning message does not inform the user to who that message is being sent.

The next chapter discusses the system's architecture, and methodology with block diagram of system.

Chapter 3

SYSTEM'S PROPOSED ARCHITECTURE

This chapter states the problem definition, methodology, and block diagram of the system.

3.1 PROBLEM DEFINITION :

Cyberbullying is intentional actions performed by an individual or a group of people via digital communication methods such as sending messages and posting comments against a victim. The System will detect cyberbullying in web chat application by using a Support Vector machine [1].

3.2 METHODOLOGY :

1. User Interaction:
    - The user can interact with the system using a Graphical User Interface (GUI).
    - The user is required to enter their Username and the name of the chatroom they wish to join.
2. Entering the Chatroom:
    - After entering the valid details, the user gains access to the specified chatroom.
    - The user can engage in conversations with other members present in the chatroom.
3. Toxicity Detection:
    - The system incorporates a machine learning model to detect the toxicity of messages within the chatroom.
    - The system analyzes each message to determine if it contains any elements of cyberbullying.

- If a message is identified as cyberbullying, the system alerts the user and provides information about the type of cyberbullying involved.

By following this methodology, users can effectively interact with the chat application, join specific chatrooms, and be alerted about any instances of cyberbullying detected within the messages.

3.3 BLOCK DIAGRAM

Supervised classification machine learning algorithms, such as Support vector Machine and TfidfVectorizer Technique, are being used to automatically detect cyberbullying in live chat. The dominance of the SVM algorithm and TfidfVectorizer Technique is that they calculate the probabilities for each class.

To apply cyberbullying detection model, a conceptual framework from [2] is referred, shown in Figure 3.1

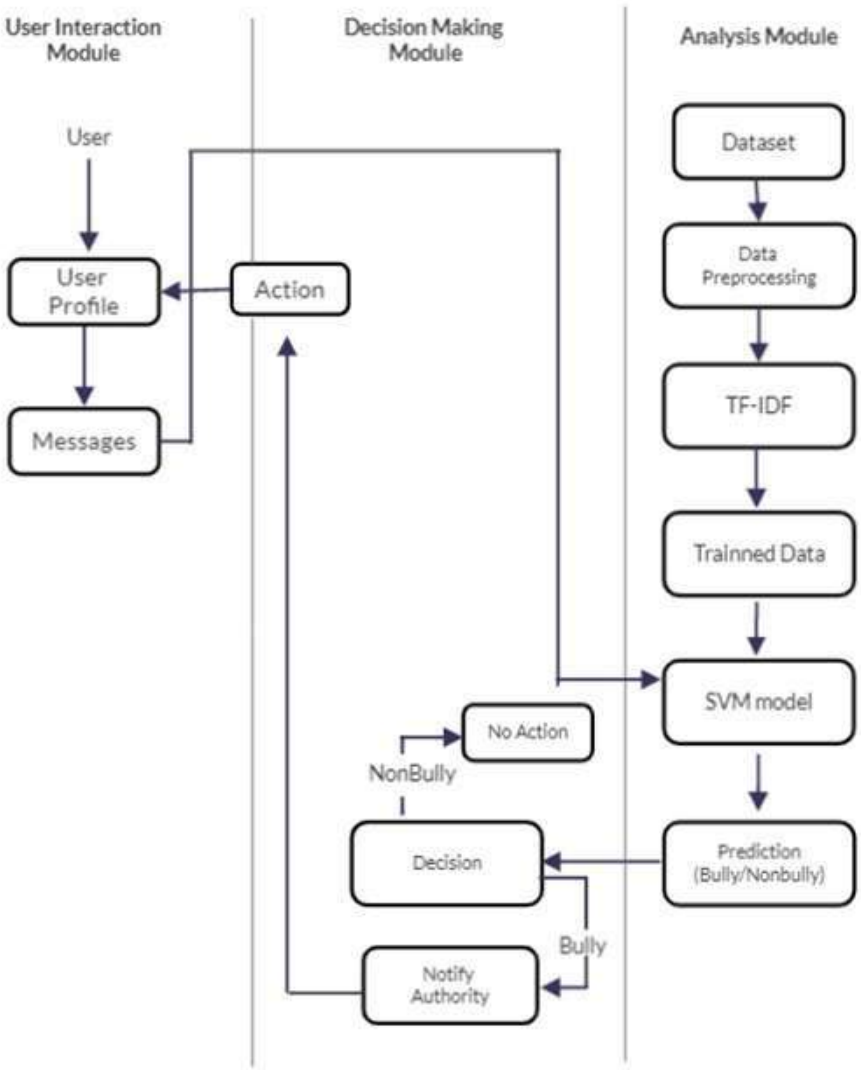

Figure 3.1: Block Diagram

This project includes the following modules:

1. User Interaction Module:

   The User Interaction Module facilitates the interaction between the user and the system. It is the user interface of the chat application that allows users to chat with others in a chatroom. It consists of the User Profile and Messaging functions.

2. Analysis Module:

   This module involves collecting the dataset, performing data pre-processing, and converting it using a vectorizer technique. The testing and training datasets are divided, and the algorithm is initialized. The features and labels are then fitted into the algorithm, and the model is saved to the system after being predicted with accuracy. The messages from the chatroom are passed to the saved trained SVM model for classification as a bully or not.

3. Decision-Making Module:

This module involves alerting the user if they are bullying others. Based on the prediction by the machine learning model, a chat message is classified as bullying or not, and a warning is displayed to the user. If the chat text is classified as not bullying, then no need to send an alert. The warning message is as follows: Warning: This message may be bullying! Category: Age Stay respectful and kind. Cyberbullying will not be tolerated.

Chapter 4

SPECIFIC REQUIREMENTS

This chapter specifies the Hardware and Software requirements of the project with the required dataset.

4.1 HARDWARE REQUIREMENT

- Computer System with RAM: 4GB or 8GB
- HDD Space: 500GB or 1TB

4.2 SOFTWARE REQUIREMENT

- Operating System
- Visual Studio Code
- Jupyter Notebook

4.3 TOOLS AND TECHNOLOGIES USED

- **Django Web Framework:**

  The core framework used for developing the web chat application is Django. Django provides a solid foundation for building web applications, offering features such as routing, request handling, authentication, and session management. It allows for easy data handling, model deployment, and user interface design.

- **Python Programming Language:**

  Python is the primary programming language used in developing the web chat application. Python's simplicity, readability, and extensive libraries made it an ideal choice for implementing various functionalities, including data processing, machine learning, and integration with Django.

- **Natural Language Processing (NLP) Libraries:**

To analyze and process the textual content of chat messages, we utilized NLP libraries in Python. Some of the key NLP libraries we used include NLTK (Natural Language Toolkit), and scikit-learn. These libraries provided a wide range of functionalities, such as tokenization, text preprocessing, feature extraction, and text classification.

- **Support Vector Machine (SVM):**

  We have employed the Support Vector Machine algorithm, available in scikit-learn, for text classification. SVM is a powerful machine learning algorithm that can effectively classify text data based on learned patterns and features. It played a crucial role in determining whether a chat message was classified as cyberbullying or not.

- **HTML, CSS, and JavaScript:**

  To create an engaging and user-friendly interface, web technologies such as HTML, CSS, and JavaScript are used. These front-end technologies are used to design and implement the chat interface, handle user interactions, and display real-time feedback on the classification results.

4.4 DATASET USED

- Cyberbullying Classification Dataset from Kaggle

As this chapter has given the overall view of the various system-specific requirement, tools, and technologies used for developing the system, the next chapter illustrates the high-level design of the system.

Chapter 5

HIGH-LEVEL DESIGN OF PROJECT UML DIAGRAM :

This chapter gives the high-level design of the project i.e. UML diagrams: Use Case diagram, Activity diagram, and Sequence diagram.

5.1 USE CASE DIAGRAM

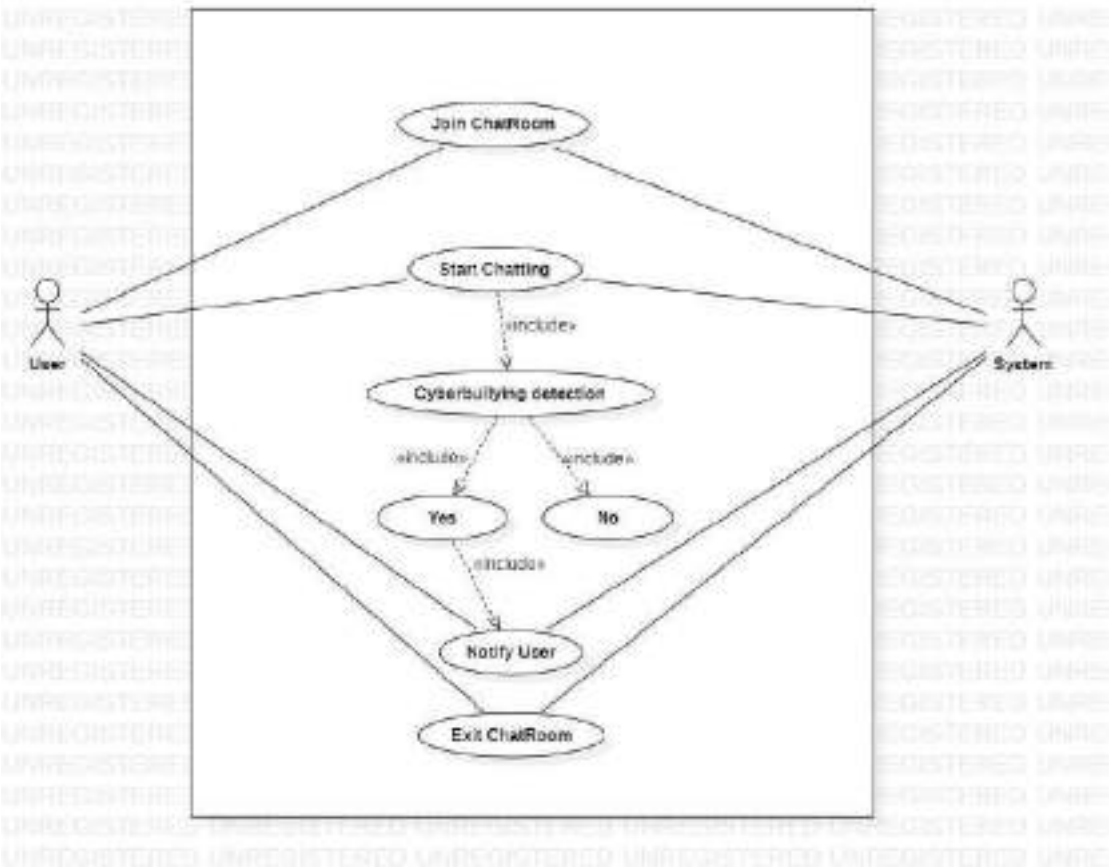

Figure 5.1: Use Case Diagram

The use case diagram illustrates the application's usage flow for the two actors: The user and System. The User initiates the process by logging into the system us-

ing their credentials. Upon successful login, the User can join a chatroom to engage with other members. Real-time messages from the chatroom are then passed on to the System for further processing and check whether the messages contain cyberbullying content or not. If a message contains cyberbullying, then the System alerts the user that stop cyberbullying.

5.2 ACTIVITY DIAGRAM

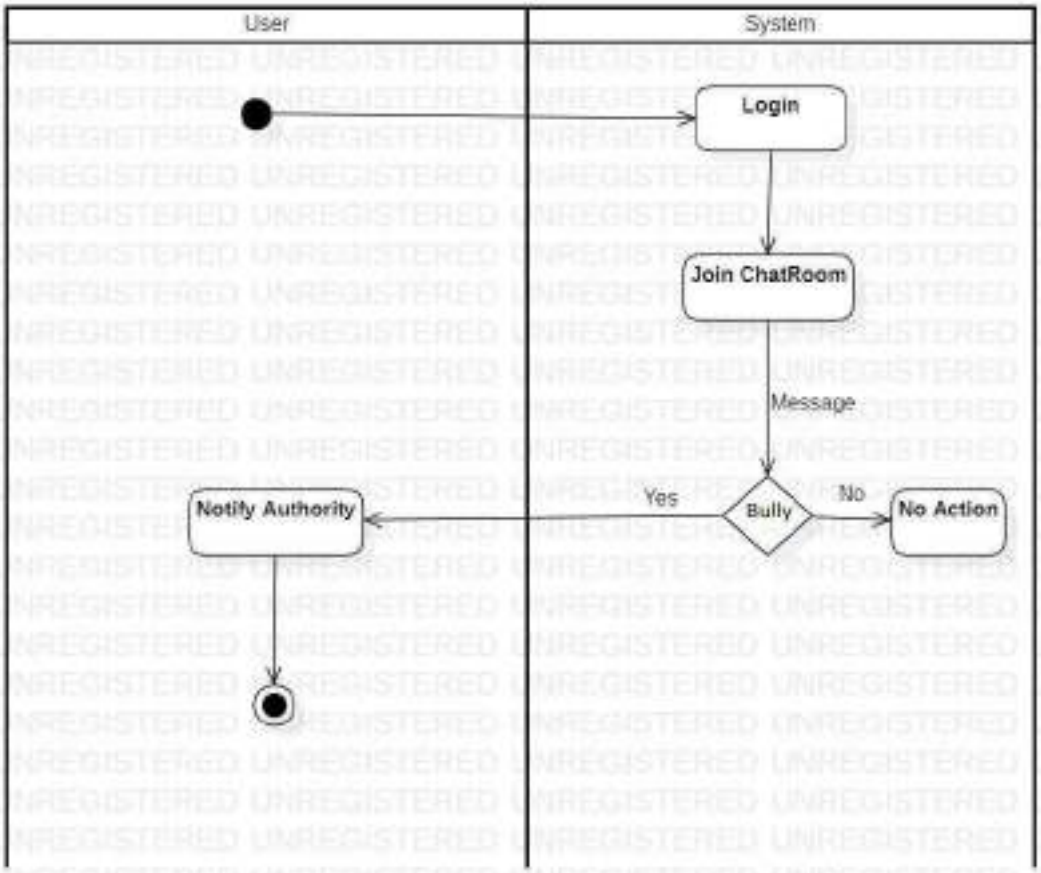

Figure 5.2: Activity Diagram

The activity diagram shows the activity phase that is happening in the application. There are two attributes of the application User and System. To start using the application, the user has to provide the credentials like username and name of the chatroom. Then the user will join the chatroom. The real-time messages from chat will get passed to the system. It will check if the message is cyberbullying or not, if the message is bullied, the system will warn the user.

5.3 SEQUENCE DIAGRAM

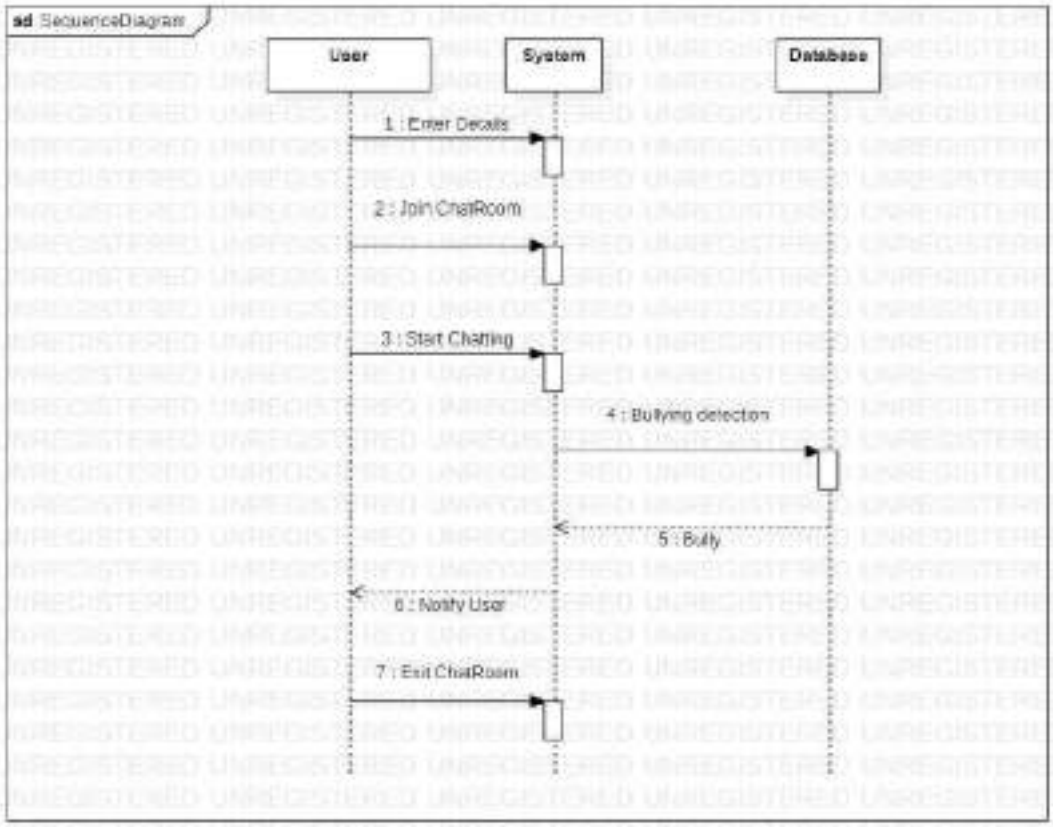

Figure 5.3: Sequence Diagram

The sequence diagram shows the sequence of actions/activities that are happening in the application. The User is an actor in the system. The user interacts with the system by using GUI. To start chatting the user has to provide credentials such as username and name of the chatroom he/she wants to join, After that he/she will be able to join the chatroom and start a conversation with group members. The real-time messages from chat will get passed to the system. It will check if the message is cyberbullying or not, if the message is bullied, the system will notify the user.

The UML diagrams were discussed in this chapter which gives the high-level design of the project. In the next chapter project flow, information on the dataset, performance parameters of the machine learning model, and efficiency issues of the system are discussed.

Chapter 6

EXPERIMENTAL SETUP / SIMULATION

This chapter focuses on the experimental setup by giving the project flow, information about the dataset being used performance parameters, and efficiency issues.

6.1 PROJECT FLOW

This project will be developed using Python, ML, and web technology.

1. The cyberbullying classification dataset is taken from Kaggle which is used to train the model.
2. The data from the dataset is preprocessed, then the cleaned data is passed to the SVM model for training purposes. The SVM model trains the data and saves the model, hence there is no need to train the model again.
3. The real-time message from the chat application is passed to the function which generates the result based on custom inputs, and alerts the user if he/she is bullying others. The message is preprocessed before passing it to that function.
4. For the frontend purpose the chat application is developed with the Django web framework. The users can join the chatroom by using the credentials like username and the name of the chatroom. If the credentials are valid the user will be redirected to the chatroom he/she wants to enter. Now the user will be able to join the group or chatroom and start chatting with other members of the chatroom.
5. If the message is cyberbullying or offensive it will alert the user to stop cyberbullying.

6.2 DATASET

About the dataset

The dataset used for this project consists of more than 47,000 tweets that have been labeled according to the class of cyberbullying. It provides valuable insights into the prevalence and impact of cyberbullying in the context of social media usage. The dataset includes information related to various categories, such as age, ethnicity, gender, religion, and other types of cyberbullying.

The collection of this dataset was motivated by the increasing usage of social media platforms across all age groups and the corresponding rise in cyberbullying incidents. Furthermore, the dataset takes into account the unique challenges posed by the COVID-19 pandemic, such as widespread school closures, increased screen time, and reduced face-to-face social interaction, which have further amplified the risk of cyberbullying. The statistics derived from the dataset are quite concerning. According to the data, approximately 36.5% of middle and high school students have reported experiencing cyberbullying, while a staggering 87% have observed instances of cyberbullying. These incidents have been linked to adverse effects on mental health, academic performance, and even thoughts of self-harm.

The utilization of this dataset within the project enabled a comprehensive analysis of the factors contributing to cyberbullying and provided insights into potential strategies for prevention and intervention.

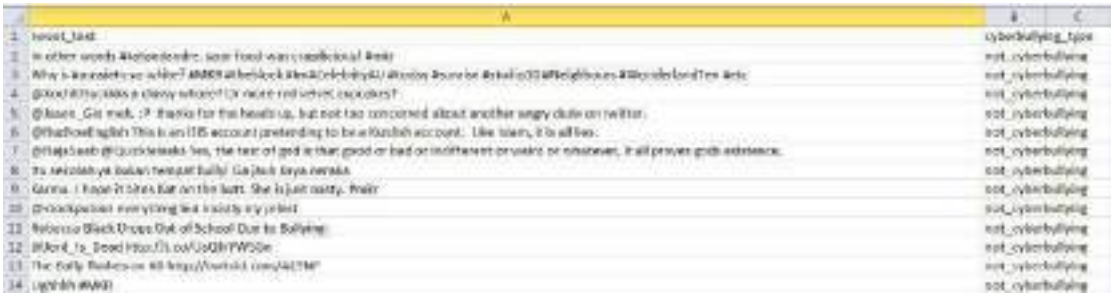

Figure 6.1: Dataset

6.3 PERFORMANCE PARAMETER

Accuracy

The accuracy of a machine learning classification algorithm is one way to measure how often the algorithm classifies a data point correctly [9].

To evaluate the effectiveness of our approach, we employed the Support Vector Machine (SVM) algorithm to train a machine learning model on the dataset. The SVM model was utilized to classify instances of cyberbullying based on various features such as age, ethnicity, gender, religion, and other relevant factors.

The Accuracy of Cyberbullying Classification Dataset using SVM is 83%. This accuracy rate reflects the percentage of correctly classified instances of cyberbullying, which demonstrates the model's ability to discern between different classes of cyberbullying accurately. The 83% accuracy rate obtained by the SVM model indicates its efficacy in classifying instances of cyberbullying within the given dataset.

Figure 6.2: Accuracy

6.4 EFFICIENCY ISSUES

1. **HUMAN DATA CHARACTERISTICS**

   In [1], Human behavior is dynamic. Knowing when online users change their way of committing cyberbullying is an important component in updating the prediction model with such changes. The Model will be able to predict the cyberbullying words which are there in the dataset.

2. **LANGUAGE DYNAMICS**

   Language is changing, and the new style of speaking is being evolved very fastly, particularly among the young generation. New slang is regularly integrated into the language culture. If the user will use another word other than the dataset, the model will not be able to recognize it.

Overall, the experimental setup is discussed in this chapter. The next chapter gives an overview of the result and evaluation metrics that are used to evaluate the system.

Chapter 7

RESULTS AND EVALUATION

Experimental results, Test cases, Working Modules of the project, and cost analysis for the development of the system are discussed in this chapter.

7.1 EXPERIMENTAL RESULTS

The project aims to promote peace and contribute to society through the use of machine learning, a trending, and emerging technology. The system is designed to detect and prevent abusive conversations during live chats. To achieve this, a Support Vector Machine algorithm is used to detect abusive words. The combination of machine learning and Python is used to train and test the model, resulting in high accuracy. The model includes features that can identify abusive words and categorize them in various types such as age, gender, ethnicity, religion, etc., and then effectively send the alert to the person and prevent cyberbullying.

The below two points shows the experimental results of the model in the form of classification report and classification report metrics.

Classification Report

Figure 7.1 shows the classification report of the machine learning model where 0 to 5 are the cyberbullying type which is encoded from the types of cyberbullying such as age, ethnicity, gender, religion, other cyberbullying, and not cyberbullying. The accuracy of the SVM model is 83%. The classification report also shows the precision, recall, and f1-score values for each cyberbullying type.

	precision	recall	f1-score	support
0	0.94	0.99	0.96	2356
1	0.97	0.99	0.98	2466
2	0.89	0.87	0.88	2393
3	0.64	0.53	0.58	2402
4	0.61	0.67	0.64	2358
5	0.94	0.97	0.95	2333
accuracy			0.83	14308
macro avg	0.83	0.83	0.83	14308
weighted avg	0.83	0.83	0.83	14308

Figure 7.1: Classification Report

Classification Report Metrics

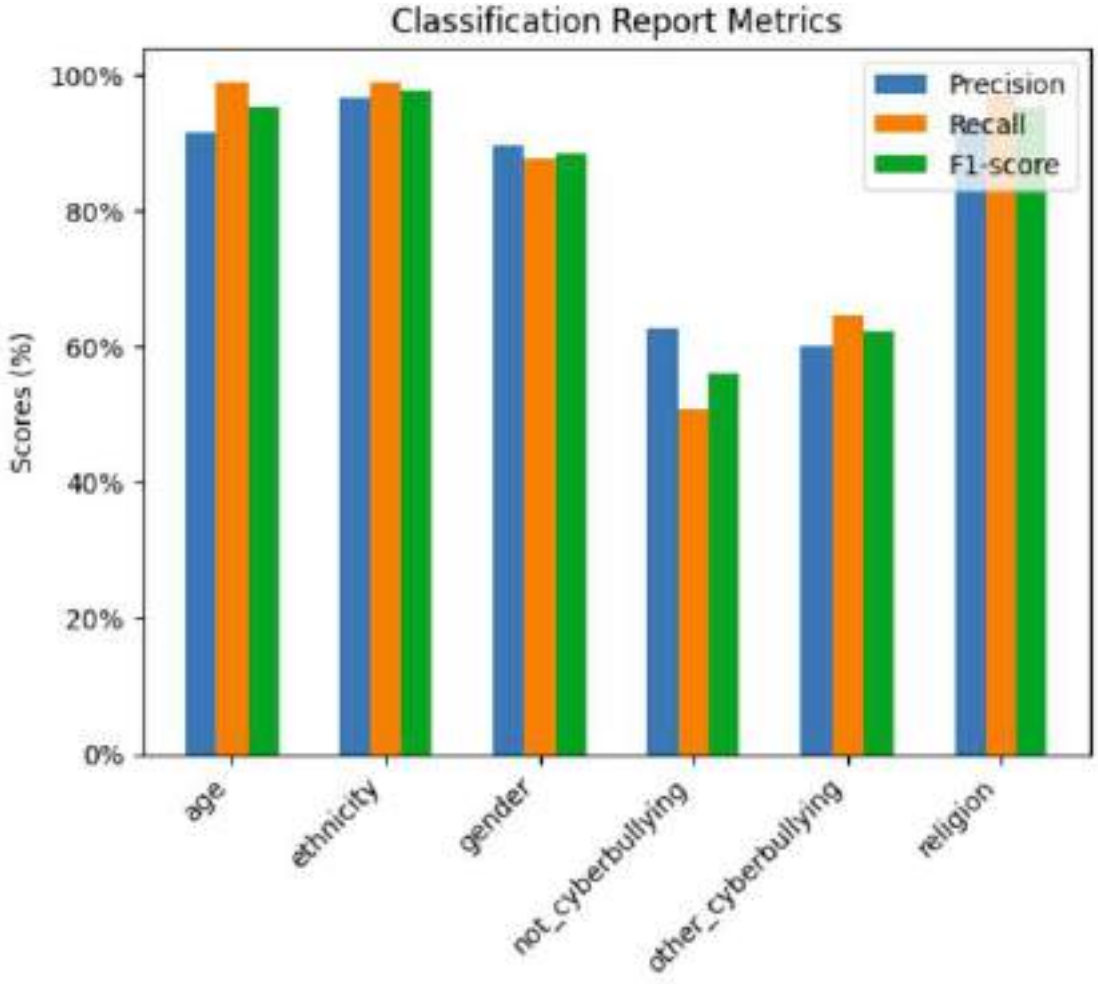

Figure 7.2: Classification Report Metrics

The bar plot in Figure 7.2, shows the classification report metrics, including precision, recall, and F1-score, for each label. The x-axis represents the different classification labels, such as age, ethnicity, gender, not_cyberbullying, other_cyberbullying, and religion. The y-axis represents the scores in percentage (%), indicating the performance of the model.

The blue bars indicate the precision scores, which measure the proportion of correctly predicted positive instances for each label. The orange bars represent the recall scores, which measure the proportion of correctly predicted positive instances out of all actual positive instances for each label. The green bars represent the F1-scores, which provide a balanced measure of precision and recall.

From the graph, It can be observed that there is varying performance of the model across different labels. The model achieves higher precision for certain labels, such as 'age' and 'gender', indicating accurate predictions of positive instances. On the other hand, labels like 'ethnicity' and 'religion' show lower precision scores, suggesting a higher proportion of false positive predictions.

In terms of recall, the model demonstrates higher performance for labels like 'gender' and 'not_cyberbullying', achieving a greater ability to identify actual positive instances. However, labels such as 'ethnicity' and 'other_cyberbullying' exhibit lower recall scores, indicating a higher proportion of false negatives.

7.2 TEST CASES

Table 7.1 shows the test cases with the expected and actual results and status if that test case has passed or failed.

Table 7.1: Test Cases

Sr. No	Test Cases	Expected Result	Actual Result	Status (Pass/Fail)
1	User can join the chatroom	User should be able to join the chatroom if the credentials are valid.	User can join the chatroom if credentials are valid.	Pass
2	System is alerting the user	If the message sent by the user is bullied, then the system should alert the user.	System alerts the user if he/she is bullying others.	Pass
3	Valid Non-Cyberbullying Message	The message is identified as non-cyberbullying and categorized accordingly.	If the message is non-cyberbullying sometimes the system categorizes it into other_cyberbullying type.	Fail
4	Valid Cyberbullying Message	The message is identified as cyberbullying and categorized into relevant types.	If the message is cyberbullying it categorizes it into relevant types.	Pass

7.3 WORKING MODULES

The current chapter points out the graphical user interface of the system. Figure 7.3 depicts the chatroom and Figure 7.4 depicts the chat containing cyberbullying and alert by the system.

7.3.1 User Interface

1. **Chatroom**

   Below figure depicts the interface of the chatroom. Multiple users can join the chatroom. There are 3 users, User 1, User 2, and User 3 in the chatroom namely "Information Technology".

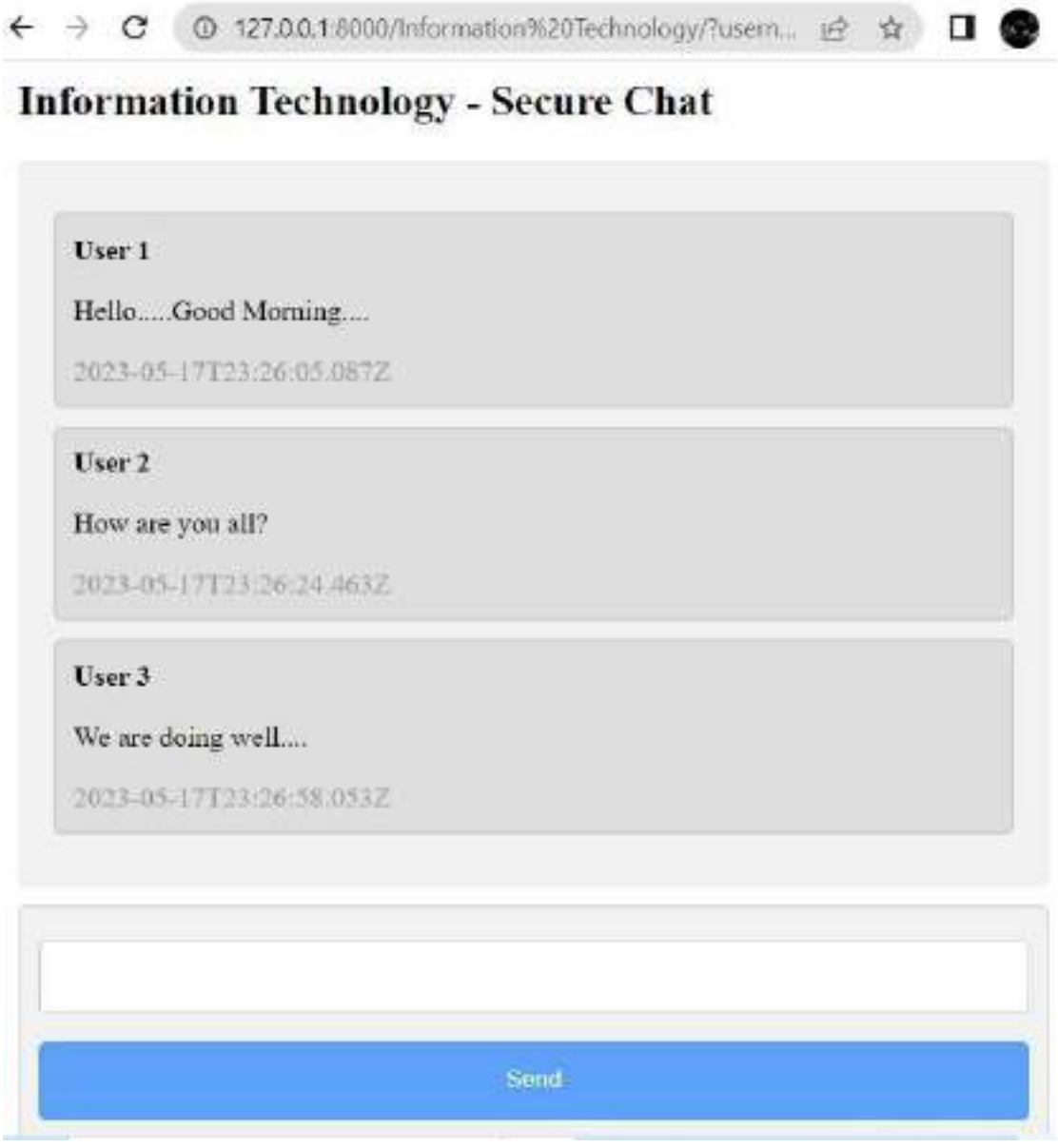

Figure 7.3: Chatroom

2. **Cyberbullying Detection**

   Below figure depicts the interface of the chatroom where User 2 has sent a message which is bullied, hence the system alerts the user and gives the type of cyberbullying.

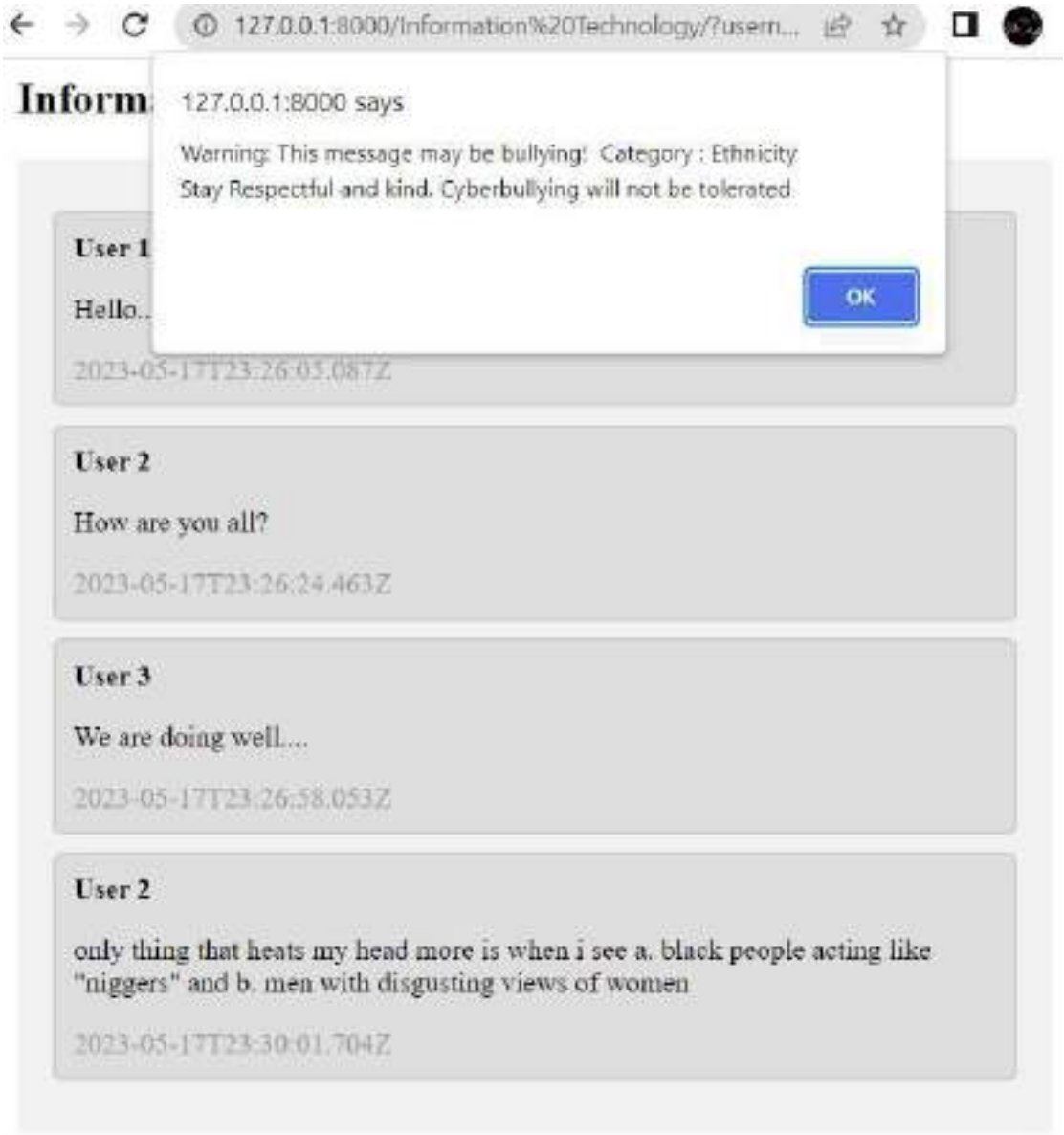

Figure 7.4: Cyberbullying Detection

7.3.2 Django Administration

Figure 7.5 and Figure 7.6 shows the Django administration interface. One of the key features of the Django web framework that is utilized in our web chat application is Django administration. Django administration provides a built-in, customizable administrative interface that allows administrators to manage various aspects of the application.

1. Message Object

    - The administration interface allowed administrators to access a list of messages in a particular chat room. They could browse through the messages to review their content, timestamp, and other relevant information. This functionality provided an overview of the conversations taking place within the chat application.

Figure 7.5 shows the message object from the Django Administration interface.

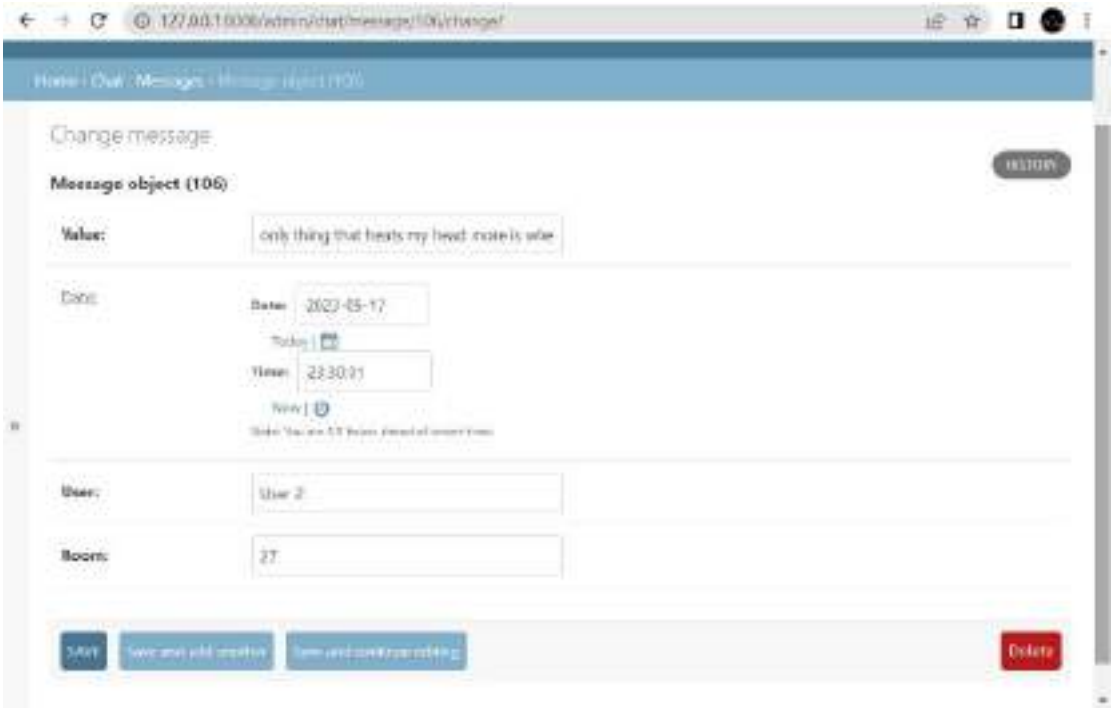

Figure 7.5: Message Object

2. Room Object

   The administration interface provided a dedicated section for managing the Room objects. This allowed administrators to perform the following tasks:

   - Create Rooms: Administrators could create new chat rooms directly from the administration panel. They could specify the room's name, description, and any other relevant details. This feature simplified the process of adding new chat rooms to the application.
   - Update Room Details: The administration interface allowed administrators to modify the details of existing chat rooms. They could edit the room's name, description, or any other associated attributes. This flexibility enabled administrators to keep the room information up to date.
   - Delete Rooms: Administrators could delete chat rooms when necessary. The administration panel provided a straightforward way to remove rooms that were no longer needed. This helped maintain a streamlined and organized chat environment.

   Figure 7.6 shows the room object from the Django Administration interface.

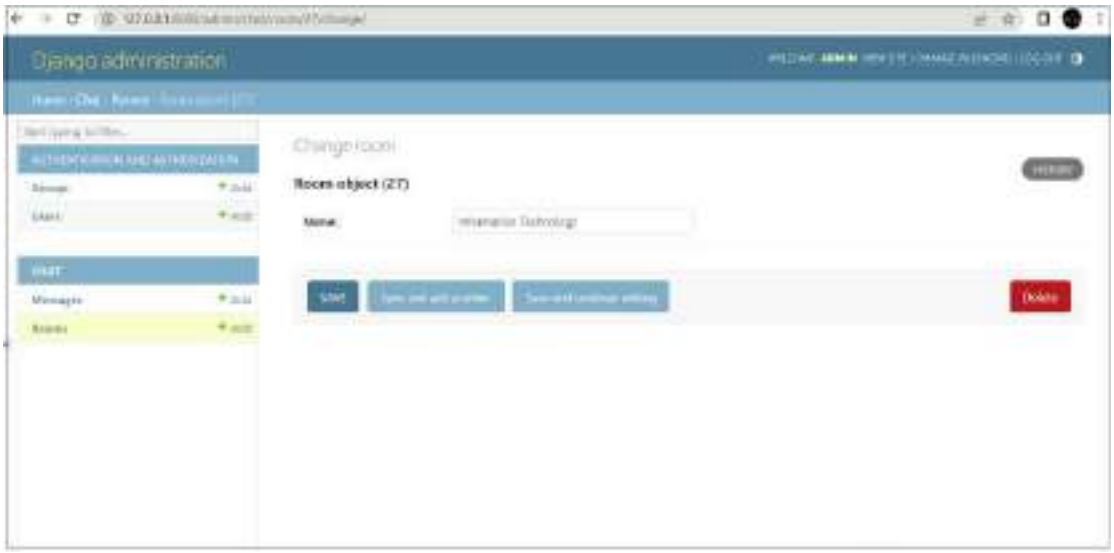

Figure 7.6: Room Object

7.4 COST ANALYSIS

COCOMO Model (COnstructive COst MOdel)

The COCOMO (COnstructive COst MOdel) is a widely used software cost estimation model developed by Barry Boehm in the late 1970s. It provides a framework for estimating the effort, cost, and schedule of software development projects based on various parameters. The model is based on the assumption that several factors influence the cost and effort required to develop software.

Semi-detached COCOMO model

The Semi-detached COCOMO model, a classification within the COCOMO framework, is used to estimate the effort, cost, and schedule of moderately complex software projects

- Predefined Values : a1=3, a2=1.12, b=2.5 , b2=0.35
  where, a1, a2, b, and b2 are the constants for each group of software products.
- Lines of Code (LOC) = 1372
- KLOC = 1.372 Where KLOc is the estimated size of the software product indicated in Kilo Lines of Code.
- Effort:

  $$\begin{aligned} \text{Effort} &= a1^*(1.374)^{a2} \\ &= 3^*(1.374)^{1.12} \\ &= 4.275 \end{aligned}$$

- Tdev: It is the estimated time to develop the software, expressed in months.

  $$\begin{aligned} \text{Tdev} &= b^*(Effort)^{b2} \\ &= 2.5^*(4.275)^{0.35} \\ &= 4.156 \ months \end{aligned}$$

- Here, We have assumed Rs.20,000 salary per engineer.
  Therefore,

  $$\begin{aligned} \text{Total Cost} &= 20000^*4.156 \\ &= 83,120 \end{aligned}$$

Total Cost = Rs.83,120

From the above illustration, if the salary per engineer is considered as Rs.20000, the total cost of the project is Rs.83120.

The next chapter gives the overview of the weekly planning and schedule of the project throughout the year.

Chapter 8

PROJECT PLANNING

This chapter gives the weekly planning of the project and the schedule of each task.

Table 8.1: Task Schedule

Sr. No.	Weeks	Tasks
1	Week 1 1 Sept to 5 Sept	Literature Survey.
2	Week 2 6 Sept to 12 Sept	Literature Survey.
3	Week 3 13 Sept to 19 Sept	Discussed two project topics and one of them was finalized.
4	Week 4 20 Sept to 26 Sept	Functionalities and approach towards the project were discussed.
5	Week 5 27 Sept to 3 Oct	Prepared Presentation on Review 1 and discussed the dataset and block diagram of the system along with the scope and objectives of the project.
6	Week 6 4 Oct to 10 Oct	Changes according to the suggestion from the Review-I presentation.
7	Week 7 11 Oct to 17 Oct	Gathering more information on detailing of project.
8	Week 8 18 Oct to 24 Oct	Identified the technology stack required to develop project.
9	Week 9 25 Oct to 31 Oct	Discussion on Experimental Setup.
10	Week 10 1 Nov to 7 Nov	Prepared the UML Diagrams.

Continued on next page

Table 8.1 – Continued from previous page

Sr. No.	Weeks	Tasks
11	Week 11 8 Nov to 14 Nov	Delivered Project Review-II. Discussed the experimental setup and UML diagrams.
12	Week 12 15 Nov to 21 Nov	Shown stage-I report to guide and made suggested changes.
13	Week 13 22 Nov to 28 Nov	Appeared for the project phase-I exam.
14	Week 14 29 Nov to 5 Dec	Submitted stage-I Report.
15	Week 15 6 Dec to 12 Dec	Started learning required technology.
16	Week 16 13 Dec to 19 Dec	Learning required technology.
17	Week 17 20 Dec to 26 Dec	Working on frontend part of chat application
18	Week 18 27 Dec to 2 Jan	Working on frontend part of chat application.
19	Week 19 3 Jan to 9 Jan	Working on frontend part of chat application.
20	Week 20 10 Jan to 16 Jan	Shown and discussed the frontend Part of the chat application.
21	Week 21 17 Jan to 23 Jan	Working on backend part of chat application.
22	Week 22 24 Jan to 30 Jan	Working on backend part of chat application.
23	Week 23 31 Jan to 6 Feb	demonstrated the chat application to guide and made some minor changes.
24	Week 24 7 Feb to 13 Feb	Discussion on Project Review-III with guide. Shown the working chat application.
25	Week 25 14 Feb to 20 Feb	Made changes in Review-III presentation suggested by guide.
26	Week 26 21 Feb to 27 Feb	Delivered project Review-III.
27	Week 27 28 Feb to 6 Mar	Made changes suggested in Review-III.

Continued on next page

Table 8.1 – Continued from previous page

Sr. No.	Weeks	Tasks
28	Week 28 7 Mar to 13 Mar	Working on machine learning module.
29	Week 29 14 Mar to 20 Mar	Working on machine learning module.
30	Week 30 21 Mar to 27 Mar	Shown the ML module to guide.
31	Week 31 28 Mar to 4 April	Review-IV Demonstrated chat application and Machine learning module.
32	Week 32 5 April to 11 April	Integrated Machine Learning module with chat application and carried out testing of the system by giving various inputs.
. 33	Week 33 12 April to 18 April	Final testing and review the overall project.
34	Week 34 19 April to 25 April	Shown and demonstrated the finalized project to guide and proceeded with Research paper publication.
36	Week 35 26 April to 2 May	Working on the project report.
36	Week 35 3 May to 10 May	Created the project report as per the guidelines.

Chapter 9

CONCLUSION

This chapter concludes the project.

The web chat application provides a safe and secure platform for users to connect with people. The machine learning model integrated with the chat application detects the toxicity of a chat. If the message is bulled, then the system sends an alert to the user with the type of cyberbullying of the message. Hence users will be able to live and maintain their healthy social life.

Bibliography

[1] Mohammed Ali Al-Garadi, Mohammad Rashid Hussain, Nawsher Khan, Ghulam Murtaza, Henry Friday Nweke, Ihsan Ali, Ghulam Mujtaba, Haruna Chiroma, Hasan Ali Khattak, AND Abdullah Gani,**Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges**, IEEE Access (Volume: 7), 22 May 2019.

[2] Shutonu Mitra, Tasfia Tasnim, Md. Arr Rafi Islam, Nafiz Imtiaz Khan, Mohammad Shahjahan Majib, **"A Framework to Detect and Prevent Cyberbullying from Social Media by Exploring Machine Learning Algorithms"**, IEEE, 10 May 2022.

[3] John Hani Mounir, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, Ammar Mohammed,**"Social Media Cyberbullying Detection using Machine Learning"**, (IJACSA) International Journal of Advanced Computer Science and Applications Vol. 10, pages 703-707, 2019.

[4] Mahamat Saleh Adoum Sanoussi, Chen Xiaohua, George K. Agordzo, Mohamed Lamine Guindo, Abdullah MMA Al Omari, Boukhari Mahamat Issa, **"Detection of Hate Speech Texts Using Machine Learning Algorithm"**, IEEE, March 2022

[5] Rahul Ramesh Dalvi, Sudhanshu Baliram Chavan, Aparna Halbe,**"Detecting A Twitter Cyberbullying Using Machine Learning'**, IEEE, June 19, 2020.

[6] D. Karthik, R. Roi, and L. Henry, **"Modeling the detection of textual cyberbullying"**, International Conference on Weblog and Social Media - Social Mobile Web Workshop, 2011.

[7] N. Vinita, L. Xue, and P. Chaoyi, **"An Effective Approach for Cyberbullying Detection"**, Communications in Information Science and Management Engineering, 2013, vol. 3, no. 5, pp.238-247.

[8] H. Homa, A. M. Sabrina, I. R. Rahat, H. Richard, L. Qin, and M. Shivakant, **"Detection of Cyberbullying Incidents on the Instagram Social Network"**, 2015.

[9] P. William, Ritik Gade, Rupesh Chaudhari, A. B. Pawar, M. A. Jawale, **"Machine Learning based Automatic Hate Speech Recognition System"**, IEEE, 27 April 20.

Appendices

A. PLAGIARISM REPORT

Document Information

Analyzed document	Project_Report_Group_14.pdf (D167592307)
Submitted	2023-05-20 07:36:00
Submitted by	rupali
Submitter email	rmbora@kkwagh.edu.in
Similarity	3%
Analysis address	rmbora.kkwagh@analysis.ouriginal.com

Sources included in the report

W	URL: https://jpinfotech.org/predicting-cyberbullying-on-social-media-in-the-big-data-era-using-mach... Fetched: 2021-02-18 11:28:58	2
W	URL: https://www.booktopia.com.au/detecting-cyberbullying-tweets-using-machine-learning-and-deep-le... Fetched: 2023-03-13 14:10:29	1
SA	**Bhavik_table of Contents-2.pdf** Document Bhavik_table of Contents-2.pdf (D165872836)	2
W	URL: https://www.sentic.net/sentire2014potha.pdf Fetched: 2020-05-24 17:34:23	1

Entire Document

Chapter 1 INTRODUCTION This chapter provides an introduction to the concept of cyberbullying with the scope and objectives of the project. The arrival of the internet has led to the widespread use of social media platforms like Facebook, Twitter, Instagram, and WhatsApp as the primary means of commu- nication. Messaging has become an essential tool in almost all sectors, including education, business, and socialization. However, it has also created new possibili- ties for harmful activities. Evidence suggests that messaging can give rise to an issue known as cyberbullying. Cyberbullying refers to the intentional use of messaging on social networking ser- vices to cause emotional, mental, or even physical harm to others. It can lead to psy- chological problems such as loneliness, low self-esteem, social anxiety, depression, and even suicide. Reports indicate that around 36% of Indian children are affected by cyberbullying. As the occurrence of cyberbullying continues to increase, it is vi- tal to study effective ways to detect and prevent it in real time. Blocking messages alone is insufficient in preventing such incidents. Instead, monitoring, processing, and analyzing text messages in real-time is necessary to make informed decisions and prevent harm to victims. Due to the aforementioned issues, several studies have been conducted to ex- plore effective techniques for detecting cyberbullying. While manual detection is the most accurate method, it is rarely employed due to the significant time and resource requirements. Therefore, the importance of developing an automatic cyberbullying detection system has increased. Although considerable research has been conducted on cyberbullying detection systems, it remains a pressing concern, and current approaches still have limitations, especially when dealing with large volumes of data. Various types of social network- ing services (SNS) can represent different forms or patterns of data. A supervised machine-learning approach can be applied to address cyberbullying from various perspectives. Our primary objective is to develop a classification model that can pre- dict text messages and prevent cyberbullying in real time. The detection process is automated, with the abusive language being quickly identified and a warning mes- sage displayed to the user who used the abusive language. 1
Cyber Bullying Detection and Prevention in Web Chat Application using SVM 1.1 SCOPE : 1. The System only focuses on the English language. 2. The system can recognize only proper and formal text. 1.2 OBJECTIVES: 1. The system will detect and prevent the cyberbullying event to deteriorate. 2. Web chat application will be used by people for safer and more secure com- munication with their friends. 3. Cyberbullying leads to an increase in stress and anger in a person, Being a vic- tim of cyberbullying also affected student's grades and hence if peoples use our system they will live a healthy social life. Dept. of Information Technology Engineering Page 2
Chapter 2 LITERATURE REVIEW This chapter discusses the literature review for the project. Cyberbullying is an act of threatening, harassing, or bullying someone through modern ways of communicating with each other and with anybody/everybody in the world via social media apps/sites. Cyberbullying is not just limited to creating a fake identity and publishing/posting some embarrassing photo or video, or unpleasant rumors about someone but also giving them threats. The impacts of cyberbullying on social media are horrifying, sometimes leading to the death of some unfortunate victims. The behavior of the victims also changes due to this, which affects their Emotions, self-confidence, and a sense of fear is also seen in such people. Al-Garadi et al. [1]

50% **MATCHING BLOCK 1/6** W

comprehensively reviewed cyberbullying prediction models and identified the main issues related to the construction of cyberbullying predic- tion models in social media. The paper provides insights into the overall process for cyberbullying detection and

mainly focuses on feature selection algorithms and the use of various machine learning algorithms for the prediction of cyberbullying behavior. The authors found that SVM was an effective and efficient algorithm for developing cyberbullying detection models. The model was trained using data con- taining cyberbullying extracted from a social network site. The paper reviews four as- pects of detecting cyberbullying messages using machine learning approaches: data collection, feature engineering, construction of cyberbullying detection models, and evaluation of constructed cyberbullying detection models. Shutonu Mitra et al. [2] The research paper proposes a conceptual framework for detecting and preventing cyberbullying on social media. The framework consists of three modules, namely the User Interaction module, the Decision-making module, and the Analysis module. The User Interaction module is responsible for collecting data from users, while the Decision-making module analyses the collected data and makes decisions based on the analysis. Finally, the Analysis module is responsible for processing the data and generating reports. The framework utilizes various machine learning algorithms to analyze and detect cyberbullying behavior on social media. The research paper aims to provide an effective solution for preventing cyberbully- ing in social media using machine learning algorithms. In recent years, several studies have been conducted on the analysis, detection, 3

Cyber Bullying Detection and Prevention in Web Chat Application using SVM and prevention of online bullying using text-mining techniques for classifying con- versations. One of these studies was conducted by John Hani et al. [3], where they proposed a classification model to detect and prevent social media bullying using neural networks and SVM. They collected their dataset from Kaggle and divided their proposed model into three major steps: 1. Pre-processing Steps: • Lowering Text • Tokenization • Stop words • Word correction 2. Feature Extraction: • Tokenization • Lowering Text 3. Classification: • SVM (Support Vector Machine) and NN classifiers were used for classifi- cation. The combination of features with an SVM classifier for predicting the class labels produced the best results. The experiments showed the models ability to detect In- sult and Hate content in text. The f1- score of the Insult and Hate class label showed a high score respectively 95.4 [4]. The Support Vector Machine achieved the highest accuracy i.e. 71.25%, while Naive Bayes achieved 52.70% accuracy. The SVM algo- rithm achieved the highest precision value i.e. 71%, while NB achieved 52% preci- sion. Also, SVM has achieved higher recall and f-score values than Naive Bayes [5]. In recent years, various studies have been conducted to explore the effective- ness of different machine-learning algorithms in detecting cyberbullying using text data from social media platforms. Karthik [6] used multi-classifiers such as Naïve Bayes, JRip, J48, and SMO with YouTube comments. Similarly, Vinita [7] employed LDA to extract features and used the weighted term frequency-inverse document fre- quency (TF-IDF) function to improve the classification with datasets from Kongre- gate, Slashdot, and MySpace. Homa [8] utilized the Support Vector Machines classi- fier with datasets from Instagram. The proposed project involves building a system using Python and Django frame- works to detect cyberbullying in live conversations. The first step is to search and download the required dataset, which will be pre-processed and used to train the model. Support Vector Machine will be used to generate the model. A web-based ap- plication will be created using the Django framework, and the generated model will be applied to live conversations to determine whether the messages are instances of cyberbullying or not. Dept. of Information Technology Engineering Page 4

Cyber Bullying Detection and Prevention in Web Chat Application using SVM 2.1 EXISTING SYSTEM This chapter gives an overview of the existing systems that are developed to detect cyberbullying. Figure 2.1 shows the interface of the "ReThink" keyboard, which is the keyboard that detects bullying words and gives an alert to the user who has typed it. Figure 2.1: Existing System "ReThink Before The Damage is Done" is an innovative and patented technology that effectively detects if the word that user has entered is cyberbullying or not and stops online hate before the damage is done. This keyboard shows the popup/alert if someone typed offensive words. It suggests clearing that message and rethink about it, but the user also has the option to choose "OK" and continue chatting. The next chapter discusses the work done related to cyberbullying issues. Dept. of Information Technology Engineering Page 5

Cyber Bullying Detection and Prevention in Web Chat Application using SVM 2.2 RELATED WORK DONE There are a few applications that are developed to detect cyberbullying, Bully Alert and the Bull Stop application are two of them which are discussed in this chapter. Bull Stop Bull Stop application is a mobile application developed for Android devices as part of a Ph.D. study at the University of Aston, United Kingdom (Salawu, 2020). The ap- plication is designed to help young people combat cyberbullying proactively, with its target audience being children aged 13 and above. The application works by using a deep learning-based algorithm to analyze messages and flag any offensive con- tent such as cyberbullying, abuse, or insults. For this application, if the application detects some offensive content in the user's social media account, it will automati- cally delete the messages to prevent them from reaching the user. Bull Stop applica- tion also allows users to block other users' social media from the application directly. This application currently only works on Twitter's social media accounts with plans for Facebook and Instagram in the future. This application's advantage is that it is easy to set up and will automatically delete the cyberbullying content just by run- ning this application in the background. Another advantage of this application will also be that this application does not allow parents to monitor their children's social media accounts to help preserve the user's social media account privacy. However, the disadvantage of this application is that it does not show the messages that are deleted, which could delete a message accidentally if the algorithm flagged the mes- sages wrongly. Another disadvantage of this application is that the user won't see the cyberbullying context as the message will be deleted before it reaches the user. Bully Alert Bully Alert application is a mobile application developed by CU Cyber Safety Lab as a research project for the University of Colorado Boulder (Lab 2018). Bully Alert ap- plication is designed for Android and can be downloaded from Google Play Store. Anyone can use this application to observe another person's social media account to get a notification from their user's profile. This application, as of now, can only be used for an Instagram account. To use this application, the user must input the user's Instagram profile name in the application. After that, the application will start observing the Instagram profile inputted into the application and then send a notifi- cation if a cyberbullying case is happening on that Instagram profile. The advantage of using this application is that it is easy to use and only requires the profile name of the person's Instagram to start observing the profile movement. The disadvantage of this application is that it can only be used if the user knows the user's Instagram profile name and the warning message does not inform the user to who that message is being sent. Dept. of Information Technology Engineering Page 6

Chapter 3 SYSTEM'S PROPOSED ARCHITECTURE This chapter states the problem definition, methodology, and block diagram of the system. 3.1 PROBLEM DEFINITION : Cyberbullying is intentional actions performed by an individual or a group of peo- ple via digital communication methods such as sending messages and posting com- ments against a victim. The System will detect cyberbullying in web chat application by using a Support Vector machine [1]. 3.2 METHODOLOGY : 1. User Interaction: • The user can interact with the system using a Graphical User Interface (GUI). • The user is required to enter their Username and the name of the chat- room they wish to join. 2. Entering the Chatroom: • After entering the valid details, the user gains access to the specified cha- troom. • The user can engage in conversations with other members present in the chatroom. 3. Toxicity Detection: • The system incorporates a machine learning model to detect the toxicity of messages within the chatroom. • The system analyzes each message to determine if it contains any ele- ments of cyberbullying. 7
Cyber Bullying Detection and Prevention in Web Chat Application using SVM • If a message is identified as cyberbullying, the system alerts the user and provides information about the type of cyberbullying involved. By following this methodology, users can effectively interact with the chat appli- cation, join specific chatrooms, and be alerted about any instances of cyberbullying detected within the messages. 3.3 BLOCK DIAGRAM Supervised classification machine learning algorithms, such as Support vector Ma- chine and TfidfVectorizer Technique, are being used to automatically detect cyber- bullying in live chat. The dominance of the SVM algorithm and TfidfVectorizer Tech- nique is that they calculate the probabilities for each class. To apply cyberbullying detection model, referred to a conceptual framework from [2], shown in Figure 3.1 Dept. of Information Technology Engineering Page 8
Cyber Bullying Detection and Prevention in Web Chat Application using SVM Figure 3.1: Block Diagram This project includes the following modules: 1. User Interaction Module: The User Interaction Module facilitates the interaction between the user and the system. It is the user interface of the chat application that allows users to chat with others in a chatroom. It consists of the User Profile and Messaging functions. 2. Analysis Module: This module involves collecting the dataset, performing data pre-processing, and converting it using a vectorizer technique. The testing and training datasets are divided, and the algorithm is initialized. The features and labels are then fitted into the algorithm, and the model is saved to the system after being pre- dicted with accuracy. The messages from the chatroom are passed to the saved trained SVM model for classification as a bully or not. 3. Decision-Making Module: Dept. of Information Technology Engineering Page 9
Cyber Bullying Detection and Prevention in Web Chat Application using SVM This module involves alerting the user if they are bullying others. Based on the prediction by the machine learning model, a chat message is classified as bul- lying or not, and a warning is displayed to the user. If the chat text is classified as not bullying, then no need to send an alert. The warning message is as fol- lows: Warning: This message may be bullying! Category: Age Stay respectful and kind. Cyberbullying will not be tolerated. Dept. of Information Technology Engineering Page 10
Chapter 4 SPECIFIC REQUIREMENTS This chapter specifies the Hardware and Software requirements of the project with the required dataset. 4.1 HARDWARE REQUIREMENT • Computer System with RAM: 4GB or 8GB • HDD Space: 500GB or 1TB 4.2 SOFTWARE REQUIREMENT • Operating System • Visual Studio Code • Jupyter Notebook 4.3 TOOLS AND TECHNOLOGIES USED • Django Web Framework: The core framework used for developing the web chat application is Django. Django provides a solid foundation for building web applications, offering fea- tures such as routing, request handling, authentication, and session manage- ment. It allows for easy data handling, model deployment, and user interface design. • Python Programming Language: Python is the primary programming language used in developing the web chat application. Python's simplicity, readability, and extensive libraries made it an ideal choice for implementing various functionalities, including data process- ing, machine learning, and integration with Django. • Natural Language Processing (NLP) Libraries: 11
Cyber Bullying Detection and Prevention in Web Chat Application using SVM To analyze and process the textual content of chat messages, we utilized NLP libraries in Python. Some of the key NLP libraries we used include NLTK (Nat- ural Language Toolkit), and scikit-learn. These libraries provided a wide range of functionalities, such as tokenization, text preprocessing, feature extraction, and text classification. • Support Vector Machine (SVM): We have employed the Support Vector Machine algorithm, available in scikit- learn, for text classification. SVM is a powerful machine learning algorithm that can effectively classify text data based on learned patterns and features. It played a crucial role in determining whether a chat message was classified as cyberbullying or not. • HTML, CSS, and JavaScript: To create an engaging and user-friendly interface, web technologies such as HTML, CSS, and JavaScript are used. These front-end technologies are used to design and implement the chat interface, handle user interactions, and display real-time feedback on the classification results. 4.4 DATASET USED • Cyberbullying Classification Dataset from Kaggle As this chapter has given the overall view of the various system-specific require- ment, tools, and technologies used for developing the system, the next chapter illus- trates the high-level design of the system. Dept. of Information Technology Engineering Page 12

Chapter 5 HIGH-LEVEL DESIGN OF PROJECT UML DIAGRAM : This chapter gives the high-level design of the project i.e. UML diagrams: Use Case diagram, Activity diagram, and Sequence diagram. 5.1 USE CASE DIAGRAM Figure 5.1: Use Case Diagram The use case diagram illustrates the application's usage flow for the two actors: The user and System. The User initiates the process by logging into the system us- 13
Cyber Bullying Detection and Prevention in Web Chat Application using SVM ing their credentials. Upon successful login, the User can join a chatroom to engage with other members. Real-time messages from the chatroom are then passed on to the System for further processing and check whether the messages contain cyber- bullying content or not. If a message contains cyberbullying, then the System alerts the user that stop cyberbullying. 5.2 ACTIVITY DIAGRAM Figure 5.2: Activity Diagram The activity diagram shows the activity phase that is happening in the applica- tion. There are two attributes of the application User and System. To start using the application, the user has to provide the credentials like username and name of the chatroom. Then the user will join the chatroom. The real-time messages from chat will get passed to the system. It will check if the message is cyberbullying or not, if the message is bullied, the system will warn the user. Dept. of Information Technology Engineering Page 14
Cyber Bullying Detection and Prevention in Web Chat Application using SVM 5.3 SEQUENCE DIAGRAM Figure 5.3: Sequence Diagram The sequence diagram shows the sequence of actions/activities that are happen- ing in the application. The User is an actor in the system. The user interacts with the system by using GUI. To start chatting the user has to provide credentials such as username and name of the chatroom he/she wants to join, After that he/she will be able to join the chatroom and start a conversation with group members. The real- time messages from chat will get passed to the system. It will check if the message is cyberbullying or not, if the message is bullied, the system will notify the user. The UML diagrams were discussed in this chapter which gives the high-level de- sign of the project. In the next chapter project flow, information on the dataset, per- formance parameters of the machine learning model, and efficiency issues of the system are discussed. Dept. of Information Technology Engineering Page 15
Chapter 6 EXPERIMENTAL SETUP / SIMULATION This chapter focuses on the experimental setup by giving the project flow, informa- tion about the dataset being used performance parameters, and efficiency issues. 6.1 PROJECT FLOW This project will be developed using Python, ML, and web technology. 1. The cyberbullying classification dataset is taken from Kaggle which is used to train the model. 2. The data from the dataset is preprocessed, then the cleaned data is passed to the SVM model for training purposes. The SVM model trains the data and saves the model, hence there is no need to train the model again. 3. The real-time message from the chat application is passed to the function which generates the result based on custom inputs, and alerts the user if he/she is bullying others. The message is preprocessed before passing it to that func- tion. 4. For the frontend purpose the chat application is developed with the Django web framework. The users can join the chatroom by using the credentials like username and the name of the chatroom. If the credentials are valid the user will be redirected to the chatroom he/she wants to enter. Now the user will be able to join the group or chatroom and start chatting with other members of the chatroom. 5. If the message is cyberbullying or offensive it will alert the user to stop cyber- bullying. 16
Cyber Bullying Detection and Prevention in Web Chat Application using SVM 6.2 DATASET About the dataset The dataset used for this project consists of more than 47,000 tweets that have been labeled according to the class of cyberbullying. It provides valuable insights into the prevalence and impact of cyberbullying in the context of social media usage. The dataset includes information related to various categories, such as age, ethnicity, gender, religion, and other types of cyberbullying. The collection of this dataset was motivated by the increasing usage of social me- dia platforms across all age groups and the corresponding rise in cyberbullying inci- dents. Furthermore, the dataset takes into account the unique challenges posed by

53%	MATCHING BLOCK 2/6	W

the COVID-19 pandemic, such as widespread school closures, increased screen time, and reduced face-to-face social interaction, which have further amplified the risk of cyberbullying.

The statistics derived from the dataset are quite concerning. Accord- ing to the data, approximately 36.5% of middle and high school students have re- ported experiencing cyberbullying, while a staggering 87% have observed instances of cyberbullying. These incidents have been linked to adverse effects on mental health, academic performance, and even thoughts of self-harm. The utilization of this dataset within the project enabled a comprehensive anal- ysis of the factors contributing to cyberbullying and provided insights into potential strategies for prevention and intervention. Figure 6.1: Dataset Dept. of Information Technology Engineering Page 17

Cyber Bullying Detection and Prevention in Web Chat Application using SVM 6.3 PERFORMANCE PARAMETER Accuracy The accuracy of a machine learning classification algorithm is one way to measure how often the algorithm classifies a data point correctly [9]. To evaluate the effectiveness of our approach, we employed the Support Vector Machine (SVM) algorithm to train a machine learning model on the dataset. The SVM model was utilized to classify instances of cyberbullying based on various fea- tures such as age, ethnicity, gender, religion, and other relevant factors. The Accuracy of Cyberbullying Classification Dataset using SVM is 83%. This ac- curacy rate reflects the percentage of correctly classified instances of cyberbullying, which demonstrates the model's ability to discern between different classes of cyber- bullying accurately. The 83% accuracy rate obtained by the SVM model indicates its efficacy in classifying instances of cyberbullying within the given dataset. Figure 6.2: Accuracy Dept. of Information Technology Engineering Page 18
Cyber Bullying Detection and Prevention in Web Chat Application using SVM 6.4 EFFICIENCY ISSUES 1. HUMAN DATA CHARACTERISTICS In [1], Human behavior is dynamic. Knowing when online users change their way of committing cyberbullying is an important component in updating the prediction model with such changes. The Model will be able to predict the cyberbullying words which are there in the dataset. 2. LANGUAGE DYNAMICS Language is changing, and the new style of speaking is being evolved very fastly, particularly among the young generation. New slang is regularly inte- grated into the language culture. If the user will use another word other than the dataset, the model will not be able to recognize it. Overall, the experimental setup is discussed in this chapter. The next chapter gives an overview of the result and evaluation metrics that are used to evaluate the system. Dept. of Information Technology Engineering Page 19
Chapter 7 RESULTS AND EVALUATION Experimental results, Test cases, Working Modules of the project, and cost analysis for the development of the system are discussed in this chapter. 7.1 EXPERIMENTAL RESULTS The project aims to promote peace and contribute to society through the use of ma- chine learning, a trending, and emerging technology. The system is designed to de- tect and prevent abusive conversations during live chats. To achieve this, a Support Vector Machine algorithm is used to detect abusive words. The combination of ma- chine learning and Python is used to train and test the model, resulting in high ac- curacy. The model includes features that can identify abusive words and categorize them in various types such as age, gender, ethnicity, religion, etc., and then effec- tively send the alert to the person and prevent cyberbullying. The below two points shows the experimental results of the model in the form of classification report and classification report metrics. Classification Report Figure 7.1 shows the classification report of the machine learning model where 0 to 5 are the cyberbullying type which is encoded from the types of cyberbullying such as age, ethnicity, gender, religion, other cyberbullying, and not cyberbullying. The ac- curacy of the SVM model is 83%. The classification report also shows the precision, recall, and f1-score values for each cyberbullying type. 20
Cyber Bullying Detection and Prevention in Web Chat Application using SVM Figure 7.1: Classification Report Classification Report Metrics Figure 7.2: Classification Report Metrics Dept. of Information Technology Engineering Page 21
Cyber Bullying Detection and Prevention in Web Chat Application using SVM The bar plot above illustrates, in Figure 7.2, the classification report metrics, in- cluding precision, recall, and F1-score, for each label. The x-axis represents the dif- ferent classification labels, such as age, ethnicity, gender, not_cyberbullying, other_cyberbullying, and religion. The y-axis represents the scores in percentage (%), indicating the per- formance of the model. The blue bars indicate the precision scores, which measure the proportion of cor- rectly predicted positive instances for each label. The orange bars represent the re- call scores, which measure the proportion of correctly predicted positive instances out of all actual positive instances for each label. The green bars represent the F1- scores, which provide a balanced measure of precision and recall. From the graph, It can be observed that there is varying performance of the model across different labels. The model achieves higher precision for certain labels, such as 'age' and 'gender', indicating accurate predictions of positive instances. On the other hand, labels like 'ethnicity' and 'religion' show lower precision scores, suggest- ing a higher proportion of false positive predictions. In terms of recall, the model demonstrates higher performance for labels like 'gender' and 'not_cyberbullying', achieving a greater ability to identify actual posi- tive instances. However, labels such as 'ethnicity' and 'other_cyberbullying' exhibit lower recall scores, indicating a higher proportion of false negatives. Dept. of Information Technology Engineering Page 22
Cyber Bullying Detection and Prevention in Web Chat Application using SVM 7.2 TEST CASES Table 7.1 shows the test cases with the expected and actual results and status if that test case has passed or failed. Table 7.1: Test Cases Sr. No Test Cases Expected Result Actual Result Status (Pass/Fail) 1 User can join the chatroom User should be able to join the chatroom if the credentials are valid. User can join the chatroom if credentials are valid. Pass 2 System is alerting the user If the message sent by the user is bullied, then the system should alert the user. System alerts the user if he/she is bullying others. Pass 3 Valid Non- Cyberbullying Message The message is identified as non-cyberbullying and categorized accordingly. If the message is non-cyberbullying sometimes the system categorizes it into other_cyberbullying type. Fail 4 Valid Cyberbullying Message The message is identified as cyberbullying and categorized into relevant types. If the message is cyberbullying it categorizes it into relevant types. Pass Dept. of Information Technology Engineering Page 23

Cyber Bullying Detection and Prevention in Web Chat Application using SVM 7.3 WORKING MODULES The current chapter points out the graphical user interface of the system. Figure 7.3 depicts the chatroom and Figure 7.4 depicts the chat containing cyberbullying and alert by the system. 7.3.1 User Interface 1. Chatroom Below figure depicts the interface of the chatroom. Multiple users can join the chatroom. There are 3 users, User 1, User 2, and User 3 in the chatroom namely "Information Technology". Figure 7.3: Chatroom Dept. of Information Technology Engineering Page 24
Cyber Bullying Detection and Prevention in Web Chat Application using SVM 2. Cyberbullying Detection Below figure depicts the interface of the chatroom where User 2 has sent a mes- sage which is bullied, hence the system alerts the user and gives the type of cyberbullying. Figure 7.4: Cyberbullying Detection Dept. of Information Technology Engineering Page 25
Cyber Bullying Detection and Prevention in Web Chat Application using SVM 7.3.2 Django Administration Figure 7.5 and Figure 7.6 shows the Django administration interface. One of the key features of the Django web framework that is utilized in our web chat application is Django administration. Django administration provides a built-in, customizable administrative interface that allows administrators to manage various aspects of the application. 1. Message Object • The administration interface allowed administrators to access a list of messages in a particular chat room. They could browse through the mes- sages to review their content, timestamp, and other relevant information. This functionality provided an overview of the conversations taking place within the chat application. Figure 7.5 shows the message object from the Django Administration interface. Figure 7.5: Message Object Dept. of Information Technology Engineering Page 26
Cyber Bullying Detection and Prevention in Web Chat Application using SVM 2. Room Object The administration interface provided a dedicated section for managing the Room objects. This allowed administrators to perform the following tasks: • Create Rooms: Administrators could create new chat rooms directly from the administration panel. They could specify the room's name, descrip- tion, and any other relevant details. This feature simplified the process of adding new chat rooms to the application. • Update Room Details: The administration interface allowed administra- tors to modify the details of existing chat rooms. They could edit the room's name, description, or any other associated attributes. This flex- ibility enabled administrators to keep the room information up to date. • Delete Rooms: Administrators could delete chat rooms when necessary. The administration panel provided a straightforward way to remove rooms that were no longer needed. This helped maintain a streamlined and or- ganized chat environment. Figure 7.6 shows the room object from the Django Administration interface. Figure 7.6: Room Object Dept. of Information Technology Engineering Page 27
Cyber Bullying Detection and Prevention in Web Chat Application using SVM 7.4 COST ANALYSIS COCOMO Model (COnstructive COst MOdel) The COCOMO (COnstructive COst MOdel) is a widely used software cost estimation model developed by Barry Boehm in the late 1970s. It provides a framework for es- timating the effort, cost, and schedule of software development projects based on various parameters. The model is based on the assumption that several factors in- fluence the cost and effort required to develop software. Semi-detached COCOMO model The Semi-detached COCOMO model, a classification within the COCOMO frame- work, is used to estimate the effort, cost, and schedule of moderately complex soft- ware projects • Predefined Values : a1=3, a2=1.12, b=2.5 , b2=0.35 where, a1, a2, b, and b2 are the constants for each group of software products. • Lines of Code (LOC) = 1372 • KLOC = 1.372 Where KLOc is the estimated size of the software product indi- cated in Kilo Lines of Code. • Effort: Effort = a1 * (1.374) a2 = 3 * (1.374) 1.12 = 4.275 • Tdev: It is the estimated time to develop the software, expressed in months. Tdev = b * (E f f or t) b2 = 2.5 * (4.275) 0.35 = 4.156 months • Here, We have assumed Rs.20,000 salary per engineer. Therefore, Total Cost = 20000 * 4.156 = 83,120 Total Cost = Rs.83,120 From the above illustration, if the salary per engineer is considered as Rs.20000, the total cost of the project is Rs.83120. The next chapter gives the overview of the weekly planning and schedule of the project throughout the year. Dept. of Information Technology Engineering Page 28
Chapter 8 PROJECT PLANNING This chapter gives the weekly planning of the project and the schedule of each task. Table 8.1: Task Schedule Sr. No. Weeks Tasks 1 Week 1 1 Sept to 5 Sept Literature Survey. 2 Week 2 6 Sept to 12 Sept Literature Survey. 3 Week 3 13 Sept to 19 Sept Discussed two project topics and one of them was finalized. 4 Week 4 20 Sept to 26 Sept Functionalities and approach towards the project were discussed. 5 Week 5 27 Sept to 3 Oct Prepared Presentation on Review 1 and discussed the dataset and block diagram of the system along with the scope and objectives of the project. 6 Week 6 4 Oct to 10 Oct Changes according to the suggestion from the Review-I presentation. 7 Week 7 11 Oct to 17 Oct Gathering more information on detailing of project. 8 Week 8 18 Oct to 24 Oct Identified the technology stack required to develop project. 9 Week 9 25 Oct to 31 Oct Discussion on Experimental Setup. 10 Week 10 1 Nov to 7 Nov Prepared the UML Diagrams. Continued on next page 29

Cyber Bullying Detection and Prevention in Web Chat Application using SVM Table 8.1 – Continued from previous page Sr. No. Weeks Tasks 11 Week 11 8 Nov to 14 Nov Delivered Project Review-II. Discussed the experimental setup and UML diagrams. 12 Week 12 15 Nov to 21 Nov Shown stage-I report to guide and made suggested changes. 13 Week 13 22 Nov to 28 Nov Appeared for the project phase-I exam. 14 Week 14 29 Nov to 5 Dec Submitted stage-I Report. 15 Week 15 6 Dec to 12 Dec Started learning required technology. 16 Week 16 13 Dec to 19 Dec Learning required technology. 17 Week 17 20 Dec to 26 Dec Working on frontend part of chat application 18 Week 18 27 Dec to 2 Jan Working on frontend part of chat application. 19 Week 19 3 Jan to 9 Jan Working on frontend part of chat application. 20 Week 20 10 Jan to 16 Jan Shown and discussed the frontend Part of the chat application. 21 Week 21 17 Jan to 23 Jan Working on backend part of chat application. 22 Week 22 24 Jan to 30 Jan Working on backend part of chat application. 23 Week 23 31 Jan to 6 Feb demonstrated the chat application to guide and made some minor changes. 24 Week 24 7 Feb to 13 Feb Discussion on Project Review-III with guide. Shown the working chat application. 25 Week 25 14 Feb to 20 Feb Made changes in Review-III presentation suggested by guide. 26 Week 26 21 Feb to 27 Feb Delivered project Review-III. 27 Week 27 28 Feb to 6 Mar Made changes suggested in Review-III. Continued on next page Dept. of Information Technology Engineering Page 30
Cyber Bullying Detection and Prevention in Web Chat Application using SVM Table 8.1 – Continued from previous page Sr. No. Weeks Tasks 28 Week 28 7 Mar to 13 Mar Working on machine learning module. 29 Week 29 14 Mar to 20 Mar Working on machine learning module. 30 Week 30 21 Mar to 27 Mar Shown the ML module to guide. 31 Week 31 28 Mar to 4 April Review-IV Demonstrated chat application and Machine learning module. 32 Week 32 5 April to 11 April Integrated Machine Learning module with chat application and carried out testing of the system by giving various inputs. . 33 Week 33 12 April to 18 April Final testing and review the overall project. 34 Week 34 19 April to 25 April Shown and demonstrated the finalized project to guide and proceeded with Research paper publication. 36 Week 35 26 April to 2 May Working on the project report. 36 Week 35 3 May to 10 May Created the project report as per the guidelines. Dept. of Information Technology Engineering Page 31
Chapter 9 CONCLUSION This chapter concludes the project. The web chat application will provide a safe and secure platform for the users to connect with people. The machine learning model integrated with the chat applica- tion will detect the toxicity of a chat, if it is bad, it can flag a message and ban users. Hence users will be able to live and maintain their healthy social life. 32
Bibliography [1]

86%	MATCHING BLOCK 3/6	W

Mohammed Ali Al-Garadi, Mohammad Rashid Hussain, Nawsher Khan, Ghu- lam Murtaza, Henry Friday Nweke, Ihsan Ali, Ghulam Mujtaba, Haruna Chiroma, Hasan Ali Khattak, AND Abdullah Gani,Predicting Cyberbullying on Social Me- dia in the Big Data Era Using Machine Learning Algorithms: Review of

Litera- ture and Open Challenges, IEEE
Access (
Volume: 7), 22 May 2019. [2] Shutonu Mitra, Tasfia Tasnim, Md. Arr Rafi Islam, Nafiz Imtiaz Khan, Mohammad Shahjahan Majib, "A Framework to Detect and Prevent Cyberbullying from So- cial Media by Exploring
Machine Learning Algorithms", IEEE, 10 May 2022. [3]

90%	MATCHING BLOCK 4/6	SA	Bhavik_table of Contents-2.pdf (D165872836)

John Hani Mounir, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, Ammar Mohammed,"Social Media Cyberbullying Detection using Ma- chine Learning", (IJACSA) International Journal of Advanced Computer Science and Applications Vol. 10, pages 703-707, 2019. [4]

Mahamat Saleh Adoum Sanoussi, Chen Xiaohua, George K. Agordzo, Mohamed Lamine Guindo, Abdullah MMA Al Omari, Boukhari Mahamat Issa, "Detection of Hate Speech Texts Using Machine Learning Algorithm", IEEE, March 2022 [5]

90%	MATCHING BLOCK 6/6	SA	Bhavik_table of Contents-2.pdf (D165872836)

Rahul Ramesh Dalvi, Sudhanshu Baliram Chavan, Apama Halbe,"Detecting A Twitter Cyberbullying Using Machine Learning',

IEEE, June 19, 2020. [6] D. Karthik,

67%	MATCHING BLOCK 5/6	W

R. Roi, and L. Henry, "Modeling the detection of textual cyberbully- ing", International Conference on Weblog and Social Media - Social Mobile Web Workshop, 2011. [7]

N. Vinita, L. Xue, and P. Chaoyi, "AnEffectiveApproachforCyberbullyingDetec- tion", Communications in Information Science and Management Engineering, 2013, vol. 3, no. 5, pp.238-247. [8] H. Homa, A. M. Sabrina, I. R. Rahat, H. Richard, L. Qin, and M. Shivakant, "De- tection of Cyberbullying Incidents on the Instagram Social Network", 2015. [9] P. William, Ritik Gade, Rupesh Chaudhari, A. B. Pawar, M. A. Jawale, "Machine Learning based Automatic Hate Speech Recognition System", IEEE, 27 April 20. 33

Hit and source - focused comparison, Side by Side

Submitted text As student entered the text in the submitted document.

Matching text As the text appears in the source.

1/6	SUBMITTED TEXT	32 WORDS	50%	MATCHING TEXT	32 WORDS

comprehensively reviewed cyberbullying prediction models and identified the main issues related to the construction of cyberbullying predic- tion models in social media. The paper provides insights into the overall process for cyberbullying detection and

comprehensively review cyberbullying prediction models and identify the main issues related to the construction of cyberbullying prediction models in SM. This paper provides insights on the overall process for cyberbullying detection and

W https://jpinfotech.org/predicting-cyberbullying-on-social-media-in-the-big-data-era-using-machine ...

2/6	SUBMITTED TEXT	23 WORDS	53%	MATCHING TEXT	23 WORDS

the COVID-19 pandemic, such as widespread school closures, increased screen time, and reduced face-to-face social interaction, which have further amplified the risk of cyberbullying.

the COVID-19 pandemic due widespread school closures, increased screen time, and decreased face-to-face social interaction. The statistics of cyberbullying

W https://www.booktopia.com.au/detecting-cyberbullying-tweets-using-machine-learning-and-deep-learn ...

3/6	SUBMITTED TEXT	40 WORDS	86%	MATCHING TEXT	40 WORDS

Mohammed Ali Al-Garadi, Mohammad Rashid Hussain, Nawsher Khan, Ghu- lam Murtaza, Henry Friday Nweke, Ihsan Ali, Ghulam Mujtaba, Haruna Chiroma, Hasan Ali Khattak, AND Abdullah Gani,Predicting Cyberbullying on Social Me- dia in the Big Data Era Using Machine Learning Algorithms: Review of

MOHAMMED ALI AL-GARADI1, MOHAMMAD RASHID HUSSAIN2, NAWSHER KHAN2, GHULAM MURTAZA1,3, HENRY FRIDAY NWEKE 1, IHSAN ALI 1, GHULAM MUJTABA1,3, HARUNA CHIROMA 4, HASAN ALI KHATTAK 5, AND ABDULLAH GANI, "Predicting Cyberbullying on Social in the Big Data Era Using Machine Learning Algorithms: Review of

W https://jpinfotech.org/predicting-cyberbullying-on-social-media-in-the-big-data-era-using-machine ...

4/6	SUBMITTED TEXT	32 WORDS	90%	MATCHING TEXT	32 WORDS

John Hani Mounir, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, Ammar Mohammed,"Social Media Cyberbullying Detection using Ma- chine Learning", (IJACSA) International Journal of Advanced Computer Science and Applications Vol. 10, pages 703-707, 2019. [4]

SA Bhavik_table of Contents-2.pdf (D165872836)

5/6	SUBMITTED TEXT	24 WORDS	67%	MATCHING TEXT	24 WORDS

R. Roi, and L. Henry, "Modeling the detection of textual cyberbully- ing", International Conference on Weblog and Social Media - Social Mobile Web Workshop, 2011. [7]

R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," International Conference on Weblog and Social Media - Social Mobile Web Workshop,

W https://www.sentic.net/sentire2014potha.pdf

6/6	SUBMITTED TEXT	14 WORDS	90%	MATCHING TEXT	14 WORDS

Rahul Ramesh Dalvi, Sudhanshu Baliram Chavan, Apama Halbe,"Detecting A Twitter Cyberbullying Using Machine Learning',

SA Bhavik_table of Contents-2.pdf (D165872836)

B. BASE PAPER

IEEE Access

Received April 21, 2019, accepted May 14, 2019, date of publication May 22, 2019, date of current version June 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918354

Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges

MOHAMMED ALI AL-GARADI[1], MOHAMMAD RASHID HUSSAIN[2], NAWSHER KHAN[2], GHULAM MURTAZA[1,3], HENRY FRIDAY NWEKE[1], IHSAN ALI[1], GHULAM MUJTABA[1,3], HARUNA CHIROMA[4], HASAN ALI KHATTAK[5], AND ABDULLAH GANI[1]

[1]Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia
[2]Collage of Computer Science, King Khalid University, Abha 61421, Saudi Arabia
[3]Department of Computer Science, Sukkur IBA University, Sukkur 65203, Pakistan
[4]Department of Computer Science, Federal College of Education (Technical), Gombe 234, Nigeria
[5]Department of Computer Science, COMSATS University Islamabad, Islamabad 45000, Pakistan

Corresponding authors: Mohammed Ali Al-Garadi (mohammedali@siswa.um.edu.my), Ihsan Ali (ihsanalichd@siswa.um.edu.my), and Ghulam Mujtaba (mujtaba@iba-suk.edu.pk)

This work was supported in part by the Deanship of Scientific Research, King Khalid University, through Research Group Project under Grant R.G.P. 1/166/40, and in part by the University of Malaya Postgraduate Research under Grant PG035-2016A.

ABSTRACT Prior to the innovation of information communication technologies (ICT), social interactions evolved within small cultural boundaries such as geo spatial locations. The recent developments of communication technologies have considerably transcended the temporal and spatial limitations of traditional communications. These social technologies have created a revolution in user-generated information, online human networks, and rich human behavior-related data. However, the misuse of social technologies such as social media (SM) platforms, has introduced a new form of aggression and violence that occurs exclusively online. A new means of demonstrating aggressive behavior in SM websites are highlighted in this paper. The motivations for the construction of prediction models to fight aggressive behavior in SM are also outlined. We comprehensively review cyberbullying prediction models and identify the main issues related to the construction of cyberbullying prediction models in SM. This paper provides insights on the overall process for cyberbullying detection and most importantly overviews the methodology. Though data collection and feature engineering process has been elaborated, yet most of the emphasis is on feature selection algorithms and then using various machine learning algorithms for prediction of cyberbullying behaviors. Finally, the issues and challenges have been highlighted as well, which present new research directions for researchers to explore.

INDEX TERMS Big data, cyberbullying, cybercrime, human aggressive behavior, machine learning, online social network, social media, text classification.

I. INTRODUCTION

Machine or deep learning algorithms help researchers understand big data [1]. Abundant information on humans and their societies can be obtained in this big data era, but this acquisition was previously impossible [2]. One of the main sources of human-related data is social media (SM). By applying machine learning algorithms to SM data, we can exploit historical data to predict the future of a wide range of applications. Machine learning algorithms provide an opportunity to effectively predict and detect negative forms of human behavior, such as cyberbullying [3]. Big data analysis can uncover hidden knowledge through deep learning from raw data [1]. Big data analytics has improved several applications, and forecasting the future has even become possible through the combination of big data and machine learning algorithms [4].

An insightful analysis of data on human behavior and interaction to detect and restrain aggressive behavior involves multifaceted angles and aspects and the merging of theorems

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

and techniques from multidisciplinary and interdisciplinary fields. The accessibility of large-scale data produces new research questions, novel computational methods, interdisciplinary approaches, and outstanding opportunities to discover several vital inquiries quantitatively. However, using traditional methods (statistical methods) in this context is challenging in terms of scale and accuracy. These methods are commonly based on organized data on human behavior and small-scale human networks (traditional social networks). Applying these methods to large online social networks (OSNs) in terms of scale and extent causes several issues. On the one hand, the explosive growth of OSNs enhances and disseminates aggressive forms of behavior by providing platforms and networks to commit and propagate such behavior. On the other hand, OSNs offer important data for exploring human behavior and interaction at a large scale, and these data can be used by researchers to develop effective methods of detecting and restraining misbehavior and/or aggressive behavior. OSNs provide criminals with tools to perform aggressive actions and networks to commit misconduct. Therefore, methods that address both aspects (content and network) should be optimized to detect and restrain aggressive behavior in complex systems.

The remainder of this paper is organized as follows. Subsection I.A presents an overview of aggressive behavior in SM, and a new means in which SM websites are utilized by users to commit aggressive behavior is highlighted. I.B summarizes the motivations for constructing prediction models to combat aggressive behavior in SM. I.C highlight the importance of constructing cyberbullying prediction models. I.D, provide the methodology followed in this paper. Section 2 presents a comprehensive review of cyberbullying prediction models for SM websites from data collection to evaluation. Section 3 discusses the main issues related to the construction of cyberbullying prediction models. Research challenges, which present new research directions, are discussed in Section 4, and the paper is concluded in Section 5.

A. RISE OF AGGRESSIVE BEHAVIOR ON SM

Prior to the innovation of communication technologies, social interaction evolved within small cultural boundaries, such as locations and families [5]. The recent development of communication technologies exceptionally transcends the temporal and spatial limitations of traditional communication. In the last few years, online communication has shifted toward user-driven technologies, such as SM websites, blogs, online virtual communities, and online sharing platforms. New forms of aggression and violence emerge exclusively online [6]. The dramatic increase in negative human behavior on SM, with high increments in aggressive behavior, presents a new challenge [6], [7]. The advent of Web 2.0 technologies, including SM websites that are often accessed through mobile devices, has completely transformed functionality on the side of users [8]. SM characteristics, such as accessibility, flexibility, being free, and having well-connected social networks, provide users with liberty and flexibility to post and write on their platforms. Therefore, users can easily demonstrate aggressive behavior [9], [10]. SM websites have become dynamic social communication websites for millions of users worldwide. Data in the form of ideas, opinions, preferences, views, and discussions are spread among users rapidly through online social communication. The online interactions of SM users generate a huge volume of data that can be utilized to study human behavioral patterns [11]. SM websites also provide an exceptional opportunity to analyze patterns of social interactions among populations at a scale that is much larger than before.

Aside from renovating the means through which people are influenced, SM websites provide a place for a severe form of misbehavior among users. Online complex networks, such as SM websites, changed substantially in the last decade, and this change was stimulated by the popularity of online communication through SM websites. Online communication has become an entertainment tool, rather than serving only to communicate and interact with known and unknown users. Although SM websites provide many benefits to users, cyber criminals can use these websites to commit different types of misbehavior and/or aggressive behavior. The common forms of misbehavior and/or aggressive behavior on OSN sites include cyberbullying [3], phishing [12], spam distribution [13], malware spreading [14], and cyberbullying [15].

Users utilize SM websites to demonstrate different types of aggressive behavior. The main involvement of SM websites in aggressive behavior can be summarized in two points [9], [15].

1) [I.] OSN communication is a revolutionary trend that exploits Web 2.0. Web 2.0 has new features that allow users to create profiles and pages, which, in turn, make users active. Unlike Web 1.0 that limits users to being passive readers of content only, Web 2.0 has expanded capabilities that allow users to be active as they post and write their thoughts. SM websites have four particular features, namely, collaboration, participation, empowerment, and timeliness [16]. These characteristics enable criminals to use SM websites as a platform to commit aggressive behavior without confronting victims [9], [15]. Examples of aggressive behavior are committing cyberbullying [17]–[19] and financial fraud [20], using malicious applications [21], and implementing social engineering and phishing [12].
2) [II.] SM websites are structures that enable information exchange and dissemination. They allow users to effortlessly share information, such as messages, links, photos, and videos [22]. However, because SM websites connect billions of users, they have become delivery mechanisms for different forms of aggressive behavior at an extraordinary scale. SM websites help cybercriminals reach many users [23].

B. MOTIVATIONS FOR PREDICTING AGGRESSIVE BEHAVIOR ON SM WEBSITES

Many studies have been conducted on the contribution of machine learning algorithms to OSN content analysis in the last few years. Machine learning research has become crucial in numerous areas and successfully produced many models, tools, and algorithms for handling large amounts of data to solve real-world problems [24], [25]. Machine learning algorithms have been used extensively to analyze SM website content for spam [26]–[28], phishing [29], and cyberbullying prediction [19], [30]. Aggressive behavior includes spam propagation [13], [31]–[34], phishing [12], malware spread [14], and cyberbullying [15]. Textual cyberbullying has become the dominant aggressive behavior in SM websites because these websites give users full freedom to post on their platforms [17], [35]–[39].

SM websites contain large amounts of text and/or non-text content and other information related to aggressive behavior. In this work, a content analysis of SM websites is performed to predict aggressive behavior. Such an analysis is limited to textual OSN content for predicting cyberbullying behavior. Given that cyberbullying can be easily committed, it is considered a dangerous and fast-spreading aggressive behavior. Bullies only require willingness and a laptop or cell phone with Internet connection to perform misbehavior without confronting victims [40]. The popularity and proliferation of SM websites have increased online bullying activities. Cyberbullying in SM websites is rampant due to the structural characteristics of SM websites. Cyberbullying in traditional platforms, such as emails or phone text messages, is performed on a limited number of people. SM websites allow users to create profiles for establishing friendships and communicating with other users regardless of geographic location, thus expanding cyberbullying beyond physical location. Anonymous users may also exist on SM websites, and this has been confirmed to be a primary cause for increased aggressive user behavior [41]. Developing an effective prediction model for predicting cyberbullying is therefore of practical significance. With all these considerations, this work performs a content-based analysis for predicting textual cyberbullying on SM websites.

The motivation of this review is explained in the following section.

C. WHY CONSTRUCTING CYBERBULLYING PREDICTION MODELS IS IMPORTANT

The motivations for carrying out this review for predicting cyberbullying on SM websites are discussed as follows. Cyberbullying is a major problem [42] and has been documented as a serious national health problem [43] due to the recent growth of online communication and SM websites. Research has shown that cyberbullying exerts negative effects on the psychological and physical health and academic performance of people [44]. Studies have also shown that cyberbullying victims incur a high risk of suicidal ideation [45], [46]. Other studies [45], [46] reported an association between cyberbullying victimization and suicidal ideation risk. Consequently, developing a cyberbullying prediction model that detects aggressive behavior that is related to the security of human beings is more important than developing a prediction model for aggressive behavior related to the security of machines.

Cyberbullying can be committed anywhere and anytime. Escaping from cyberbullying is difficult because cyberbullying can reach victims anywhere and anytime. It can be committed by posting comments and statuses for a large potential audience. The victims cannot stop the spread of such activities [47]. Although SM websites have become an integral part of users' lives, a study found that SM websites are the most common platforms for cyberbullying victimization [48]. A well-known characteristic of SM websites, such as Twitter, is that they allow users to publicly express and spread their posts to a large audience while remaining anonymous [9]. The effects of public cyberbullying are worse than those of private ones, and anonymous scenarios of cyberbullying are worse than non-anonymous cases [49], [50]. Consequently, the severity of cyberbullying has increased on SM websites, which support public and anonymous scenarios of cyberbullying. These characteristics make SM websites, such as Twitter, a dangerous platform for committing cyberbullying [43].

Recent research has indicated that most experts favor the automatic monitoring of cyberbullying [51]. A study that examined 14 groups of adolescents confirmed the urgent need for automatic monitoring and prediction models for cyberbullying [52] because traditional strategies for coping with cyberbullying in the era of big data and networks do not work well. Moreover, analyzing large amounts of complex data requires machine learning-based automatic monitoring.

1) CYBERBULLYING ON SM WEBSITES

Most researchers define cyberbullying as using electronic communication technologies to bully people [53]. Cyberbullying may exist in different types or forms, such as writing aggressive posts, harassing or bullying a victim, making hateful posts, or insulting the victim [54], [55]. Given that cyberbullying can be easily committed, it is considered a dangerous and fast-spreading aggressive behavior. Bullies only require willingness and a laptop or cell phone connected to the Internet to perform misbehavior without confronting the victims [40]. The popularity and proliferation of SM websites have increased online bullying activities. Cyberbullying on SM websites is performed on a large number of users due to the structural characteristics of SM websites [48].

Cyberbullying in traditional platforms, such as emails or phone text messages, is committed on a limited number of people. SM websites allow users to create profiles for establishing friendships and interacting with other online users regardless of geographic location, thus expanding cyberbullying beyond physical location. Moreover, anonymous users may exist on SM websites, and this has been confirmed to be a primary cause of increased aggressive user behavior [41].

The nature of SM websites allows cyberbullying to occur secretly, spread rapidly, and continue easily [54]. Consequently, developing an effective prediction model for predicting cyberbullying is of practical significance. SM websites contain large amounts of text and/or non-text content and information related to aggressive behavior.

D. METHODOLOGY

This section presents the methodology used in this work for a literature search. Two phases were employed to retrieve published papers on cyberbullying prediction models. The first phase included searching for reputable academic databases and search engines. The search engines and academic databases used for the retrieval of relevant papers were as follows: Scopus, Clarivate Analytics' Web of Science, DBLP Computer Science Bibliography, ACM Digital Library, ScienceDirect, SpringerLink, and IEEE Xplore. The major keywords used for the literature search were coined in relation to social media as follows: cyberbullying, aggressive behavior, big data, and cyberbullying models. The second phase involved searching for literature through Qatar University's digital library. The articles retrieved from the search were scrutinized to ensure that the articles met the inclusion criteria. According to the inclusion criteria, for an article to be selected for the survey, it must report an empirical study describing the prediction of cyberbullying on SM sites. Otherwise, the article would be excluded in the selection. Many articles were rejected based on titles. The abstract and conclusion sections were examined to ensure that articles satisfied the screening criteria, and those that did not satisfy the criteria were excluded from the survey.

II. PREDICTING CYBERBULLYING ON SOCIAL MEDIA IN THE BIG DATA ERA USING MACHINE LEARNING ALGORITHMS

Our world is currently in the big data era because 2.5 quintillion bytes of data are generated daily [56]. Organizations continuously generate large-scale data. These large-scale datasets are generated from different sources, including the World Wide Web, social networks, and sensor networks [57]. Big data have nine characteristics, namely, volume, variety, variability and complexity, velocity, veracity, value, validity, verdict, and visibility [58]. For example, Flickr generates almost 3.6 TB of data, Google is believed to process almost 20,000 TB of data per day, and the Internet gathers an estimated 1.8 PB of data daily [59].

SM is an online platform that provides users an opportunity to create an online community, share information, and exchange content. SM users and the interaction among organizations, people, and products are responsible for the massive amount of data generated on SM platforms. SM platforms, such as Facebook, YouTube, blogs, Instagram, Wikipedia, and Twitter, are of different types. The data generated by SM outlets can be structured or unstructured in form. SM analytics is the analysis of structured and unstructured data generated by SM outlets. SM analytics can be in any of the following forms: link prediction, community, content, social influence, structured, and unstructured. SM is now in the big data era. For example, Facebook stores 260 billion photographs in over 20 PB of storage space, and up to one million pictures are processed per second. YouTube receives 100 hours of downloaded videos in each minute [60].

The most common means of constructing cyberbullying prediction models is to use a text classification approach that involves the construction of machine learning classifiers from labeled text instances [19], [38], [61]–[63]. Another means is to use a lexicon-based model that involves computing orientation for a document from the semantic orientation of words or phrases in the document [64]. Generally, the lexicon in lexicon-based models can be constructed manually (similar to the approaches used in [65]) or automatically by using seed words to expand the list of words [66]. However, cyberbullying prediction using the lexicon-based approach is rare in literature. The primary reason is that the texts on SM websites are written in an unstructured manner, thus making it difficult for the lexicon-based approach to detect cyberbullying based only on lexicons [67]–[69]. However, lexicons are used to extract features, which are often utilized as inputs to machine learning algorithms. For example, lexicon-based approaches, such as using a profane-based dictionary to detect the number of profane words in a post, are adopted as profane features to machine learning models [70]. The key to effective cyberbullying prediction is to have a set of features that are extracted and engineered [71]. Features and their combinations are crucial in the construction of effective cyberbullying prediction models [70], [71]. Most studies on cyberbullying prediction [19], [38], [62], [72], [73] used machine learning algorithms to construct cyberbullying prediction models. Machine learning-based models exhibit decent performance in cyberbullying prediction [74]. Consequently, this work reviews the construction of cyberbullying prediction models based on machine learning.

The machine learning field focuses on the development and application of computer algorithms that improve with experience [75], [76]. The objective of machine learning is to identify and define the patterns and correlations between data. The importance of analyzing big data lies in discovering hidden knowledge through deep learning from raw data [1]. Machine learning can be described as the adoption of computational models to improve machine performance by predicting and describing meaningful patterns in training data and the acquisition of knowledge from experience [77]. When this concept is applied to OSN content, the potential of machine learning lies in exploiting historical data to detect, predict, and understand large amounts of OSN data. For example, in supervised machine learning for classification application, classification is learned with the help of suitable examples from a training dataset. In the testing stage, new data are fed into the model, and instances are classified to a specified class learned during the training stage. Then, classification performance is evaluated.

This section reviews the most common processes in the construction of cyberbullying prediction models for SM websites based on machine learning. The review covers data collection, feature engineering, feature selection, and machine learning algorithms.

A. DATA COLLECTION

Data are important components of all machine learning-based prediction models. However, data (even ''Big Data'') are useless on their own until knowledge or implications are extracted from them. Data extracted from SM websites are used to select training and testing datasets. Supervised prediction models aim to provide computer techniques to enhance prediction performance in defined tasks on the basis of observed instances (labeled data) [78]. Machine learning models for a certain task primarily aim to generalize; a successful model should not be limited to examples in a training dataset only [79] but must include unlabeled real data. Data quantity is inconsequential; what is crucial is whether or not the extracted data represent activities on SM websites well [80]–[82]. The main data collection strategies in previous cyberbullying prediction studies on SM websites can be categorized into data extracted from SM websites by using either keywords, that is, words, phrases, or hashtags (e.g., [19], [43], [83]–[85]), or by using user profiles (e.g., [38], [62], [70], [86]). The issues in these data collection strategies and their effects on the performance of machine learning algorithms are highlighted in the Data Collection section (related issues).

B. FEATURE ENGINEERING

Feature is a measurable property of a task that is being observed [87]. The main purpose of engineering feature vectors is to provide machine learning algorithms with a set of learning vectors through which these algorithms learn how to discriminate between different types of classes [76]. Feature engineering is a key factor behind the success and failure of most machine learning models [79]. The success and failure of prediction may be based on several elements. The most significant element is the features used to train the model [78]. Most of the effort in constructing cyberbullying prediction models using learning algorithms is devoted to this task [61], [62], [72]. In this context, the design of the input space (i.e., features and their combinations that are provided as an input to the classifier) is vital.

Proposing a set of discriminative features, which are used as inputs to the machine learning classifier, is the main step toward constructing an effective classifier in many applications [76]. Feature sets can be created based on human-engineered observations, which rely on how features correlate with the occurrences of classes [76]. For example, recent cyberbullying studies [88]–[94] established the correlation between different variables, such as age, gender, and user personality, and cyberbullying occurrence. These observations can be engineered into a practical form (feature) to allow the classifier to discriminate between cyberbullying and non-cyberbullying and can thus be used to develop effective cyberbullying prediction models. Proposing features is an important step toward improving the discrimination power of prediction models [76], [79]. Similarly, proposing a set of significant features of cyberbullying engagement on SM websites is important in developing effective prediction models based on machine learning algorithms [68], [95].

State-of-the-art research has developed features to improve the performance of cyberbullying prediction. For example, a lexical syntactic feature has been proposed to deal with the prediction of offensive language; this method is better than traditional learning-based approaches in terms of precision [18]. Dadvar *et al.* examined gender information from profile information and developed a gender-based approach for cyberbullying prediction by using datasets from Myspace as a basis. The gender feature was selected to improve the discrimination capability of a classifier. Age and gender were included as features in other studies [17], [61], but these features are limited to the information provided by users in their online profiles.

Several studies focused on cyberbullying prediction based on profane words as a feature [35], [68], [70], [95], [96]. Similarly, a lexicon of profane words was constructed to indicate bullying, and these words were used as features for input to machine learning algorithms [97], [98]. Using profane words as features demonstrates a significant improvement in model performance. For example, the number of ''bad'' words and the density of ''bad'' words were proposed as features for input to machine learning in a previous work [70]. The study concluded that the percentage of ''bad'' words in a text is indicative of cyberbullying. Another research [85] expanded a list of pre-defined profane words and allocated different weights to create bullying features. These features were concatenated with bag-of-words and latent semantic features and used as a feature input for a machine learning algorithm.

Reference [19] proposed features, such as pronouns and skip grams, as additional features to traditional models, such as bag of words (n-gram $n = 1$). The authors claimed that adding these features improved the overall classification accuracy. Another study [62] analyzed textual cyberbullying associated with comments on images in Instagram and developed a set of features from text comprising traditional bag-of-words features, comment counts for an image, and post counts within less than one hour of posting the image. Features mined from user and media information, including the number of followers and likes, and shared media and features from image content, such as image types, were added [62]. The combination of all features improved the overall classification performance [62].

The context-based approach is better than the list-based approach in developing the feature vector [37]. However, the diversity and complexity of cyberbullying do not always support this conclusion. Several studies [68], [72], [96], [99] discussed how sentiment analysis can improve the discrimination power of a classifier to distinguish between

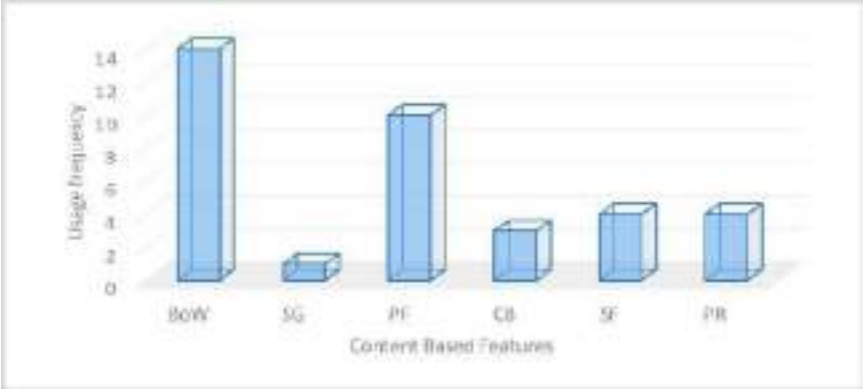

FIGURE 1. Depicting feature types used in cyberbullying prediction: Content-based features.

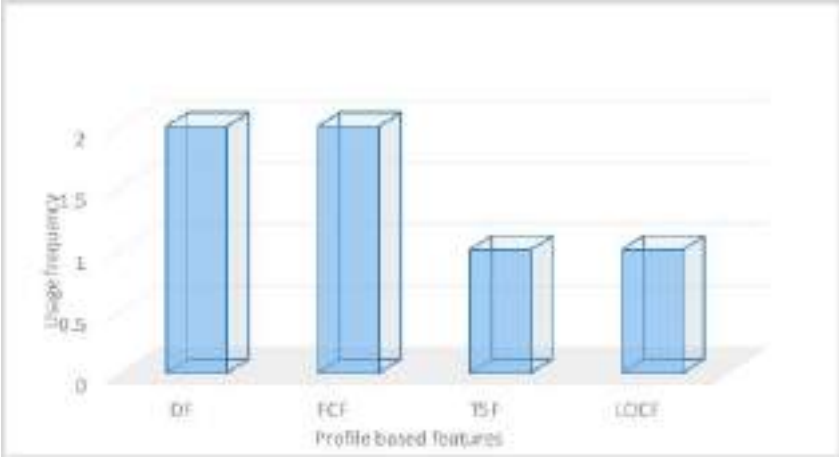

FIGURE 2. Depicting feature types used in cyberbullying prediction: Profile-based features.

cyberbullying and normal posts. These studies assumed that sentiment features are a good signal for cyberbullying occurrence. In another study that aimed to establish ways of reducing cyberbullying activities by predicting troll profiles, the researchers proposed a model to identify and associate troll profiles in Twitter; they assumed that predicting troll profiles is an important step toward predicting and stopping cyberbullying occurrence on SM websites [38]. This study proposed features based on tweeted text, posting time, language, and location to improve the identification of authorship of posts and determine whether a profile is troll or not. Reference [99] merged features from the structure of SM websites (e.g., degree, closeness, betweenness, and eigenvector centralities as well as clustering coefficient) with features from users (e.g., age and gender) and content (e.g., length and sentiment of a post). Combining these features improves the final machine learning accuracy [99]. Table 1 shows a comparison of the different features used in cyberbullying prediction literature. affect prediction performance. If the constructed features contain a large set of features that individually associate well with class, then the learning process will be effective. This condition explains why most of the discussed studies aimed to produce many features. The input features should reflect the behavior related to the occurrence of textual cyberbullying. However, the set of features should be analyzed using feature selection algorithms. Feature selection algorithms are adopted to decide which features are most probably relevant or irrelevant to classes.

C. FEATURE SELECTION ALGORITHMS

Feature selection algorithms were rarely adopted in state-of-the-art research to perform cyberbullying prediction on SM websites via machine learning (all extracted features are used to train the classifiers). Most of the examined studies (e.g., [18], [61], [68], [70]–[72], [85], [95], [96], [99]) did not use feature selection to decide which features are important in training machine learning algorithms. Two studies [19], [62] used chi-square and PCA to select a significant feature from extracted features. These feature selection algorithms are briefly discussed in following subsections.

1) INFORMATION GAIN

Information gain is the estimated decrease in entropy produced by separating examples based on specified features. Entropy is a well-known concept in information theory; it describes the (im)purity of an arbitrary collection of examples [100].

TABLE 1. Summary of feature types used in cyberbullying prediction literature.

Study	Content-based Features						Profile-based Features			
	BoW	SG	PF	CB	SF	PR	DF	FCF	TSF	LOCF
[19]	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗
[18]	✓	✗	✓	✓	✗	✓	✗	✗	✗	✗
[61]	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗
[95]	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
[72]	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗
[62]	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗
[68]	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
[74]	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
[85]	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗
[99]	✓	✗	✓	✗	✓	✓	✓	✓	✗	✗
[70]	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
[96]	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
[43]	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
[38]	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓
[71]	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗

**BoW = bag of words, SG = skip gram, PF = profanity features, SF = sentiment features, PR = pronouns, DF = demographic features (e.g., age and gender), FCF = friends or follower count features, TSF = timestamp features, LOCF = location of post feature

Information gain is used to calculate the strength or importance of features in a classification model according to the class attribute. Information gain [101] evaluates how well a specified feature divides training datasets with respect to class labels, as explained in the following equations. Given a training dataset (Tr), the entropy of (Tr) is defined as.

$$I(Tr) = -\sum P_n \log_2 P_n, \tag{1}$$

where P_n is the probability that Tr belongs to class n.

For attribute Att datasets, the expected entropy is calculated as

$$I(Att) = \sum \left(\frac{Tr_{Att}}{Tr} \right) \times I\left(Tr_{Att}\right). \tag{2}$$

The information gain of attribute Att datasets is

$$IG(Att) = I(Tr) - I(Att) \tag{3}$$

2) PEARSON CORRELATION

Correlation-based feature selection is commonly used in reducing feature dimensionality and evaluating the discrimination power of a feature in classification models. It is also a straightforward model for selecting significant features. Pearson correlation measures the relevance of a feature by computing the Pearson correlation between it and a class. The Pearson correlation coefficient measures the linear correlation between two attributes [102]. The subsequent value lies between -1 and $+1$, with -1 implying absolute negative correlation (as one attribute increases, the other decreases), $+1$ denoting absolute positive correlation (as one attribute increases, the other also increases), and 0 denoting the absence of any linear correlation between the two attributes. For two attributes or features X and Y, the Pearson correlation coefficient measures the correlation [103] as follows:

$$r_{xy} = \frac{\sum (x_i - \overline{x})\,(y_i - \overline{y})}{(n-1)\,S_x S_y}, \tag{4}$$

where $\overline{x}$ and $\overline{y}$ are the sample means for X and Y, respectively; S_x and S_y are the sample standard deviations for X and Y, respectively; and n is the size of the sample used to compute the correlation coefficient [103].

3) CHI-SQUARE TEST

Another common feature selection model is the chi-square test. This test is used in statistics, among other variables, to test the independence of two occurrences. In feature selection, chi-square is used to test whether the occurrences of a feature and class are independent. Thus, the following quantity is assumed for each feature, and they are ranked by their score.

$$N = \frac{N\left[P(f, c_i)P(\overline{f}, \overline{c}_i) - P(f, \overline{c}_i)P(\overline{f}, c_i)\right]}{P(f)P(\overline{f})P(c_i)P(\overline{c}_i)} \tag{5}$$

The chi-square test [104] assesses the independence between feature f and class c_i, in which N is the total number of documents.

D. MACHINE LEARNING ALGORITHMS

Many types of machine learning algorithms exist, but nearly all studies on cyberbullying prediction in SM websites used the most established and widely used type, that is, supervised machine learning algorithms [67], [99]. The accomplishment of machine learning algorithms is determined by the degree to which the model accurately converts various types of prior observation or knowledge about the task. Much of the practical application of machine learning considers the details

TABLE 2. Summary of machine learning algorithms tested in cyberbullying literature.

Study	SVM	NB	RF	DT	KNN	LR	ARM	RB
[19]	✓	✗	✗	✗	✗	✓	✗	✗
[18]	✓	✓	✗	✗	✗	✗	✗	✗
[61]	✓	✗	✗	✗	✗	✗	✗	✗
[95]	✓	✓	✗	✓	✗	✗	✗	✓
[38]	✓	✓	✓	✓	✓	✗	✗	✗
[86]	✗	✗	✓	✗	✗	✗	✗	✗
[72]	✓	✗	✗	✗	✗	✗	✗	✗
[62]	✓	✗	✗	✗	✗	✗	✗	✗
[74]	✗	✓	✗	✗	✗	✗	✗	✗
[73]	✓	✓	✗	✗	✗	✓	✗	✗
[84]	✗	✗	✗	✗	✗	✗	✓	✗
[71]	✓	✗	✗	✗	✗	✗	✗	✗
[85]	✓	✗	✗	✗	✗	✗	✗	✗
[99]	✓	✓	✗	✗	✗	✓	✗	✗
[70]	✓	✗	✗	✓	✗	✗	✗	✓
[96]	✓	✗	✗	✗	✗	✗	✗	✗

** SVM = support vector machine family, NB = naïve Bayes, RF = random forest, DT = decision tree family, KNN = K-nearest neighbor, LR = logistic regression, ARM = association rule mining, RB = rule-based algorithms

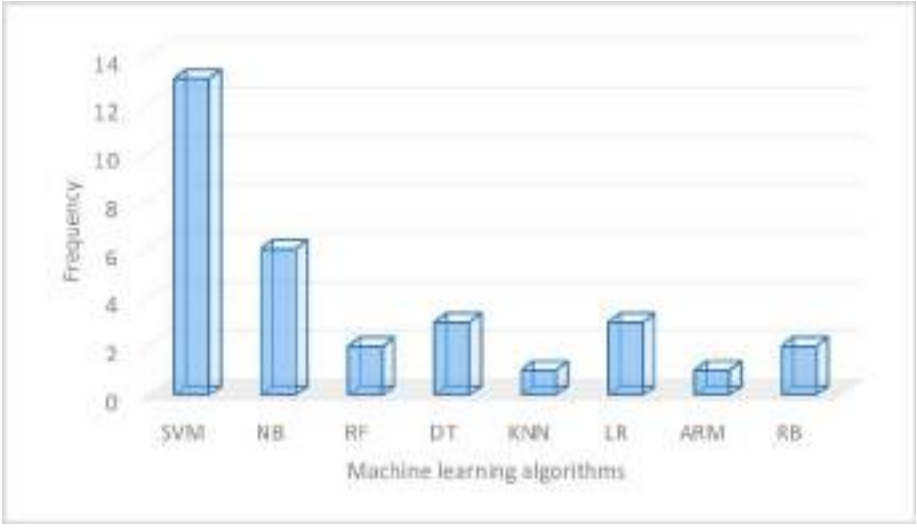

FIGURE 3. Machine learning algorithms applied in cyberbullying prediction.

of a particular problem. Then, an algorithmic model that allows for the accurate encoding of the facts is selected. However, no optimal machine learning algorithm works best for all problems [73], [105], [106]. Therefore, most researchers selected and compared many supervised classifiers to determine the ideal ones for their problem. Classifier selection is generally based on the most commonly used classifiers in the field and the data features available for experiments. However, researchers can only decide which algorithms to adopt for constructing a cyberbullying prediction model by performing a comprehensive practical experiment as a basis. Table 2 summarizes the commonly used machine learning algorithms for constructing cyberbullying prediction models.

The following sections describe the machine learning algorithms commonly used for constructing cyberbullying prediction models (Table 2).

1) SUPPORT VECTOR MACHINE IN CYBERBULLYING

Support vector machine (SVM) is a supervised machine learning classifier that is commonly used in text classification [107]. SVM is constructed by generating a separating hyperplane in the feature attributes of two classes, in which the distance between the hyperplane and the adjacent data point of each class is maximized [108]. Theoretically, SVM was developed from statistical learning theory [109]. In the SVM algorithm, the optimal separation hyperplane pertains to the separating hyperplane that minimizes misclassifications that is achieved in the training step. The approach is based on minimized classification risks [106], [110]. SVM was initially established to classify linearly separable classes. A 2D plane comprises linearly separable objects from different classes (e.g., positive or negative). SVM aims to separate

the two classes effectively. SVM identifies the exceptional hyperplane that provides the maximum margin by maximizing the distance between the hyperplane and the nearest data point of each class.

In real-time applications, precisely determining the separating hyperplane is difficult and nearly impossible in several cases. SVM was developed to adapt to these cases and can now be used as a classifier for non-separable classes. SVM is a capable classification algorithm because of its characteristics. Specifically, SVM can powerfully separate non-linearly divisible features by converting them to a high-dimensional space using the kernel model [111].

The advantage of SVM is its high speed, scalability, capability to predict intrusions in real time, and update training patterns dynamically.

SVM has been used to develop cyberbullying prediction models and found to be effective and efficient. For example, Chen *et al.* [18] applied SVM to construct a cyberbullying prediction model for the detection of offensive content in SM. SM content with potential cyberbullying were extracted, and the SVM cyberbullying prediction model was applied to detect offensive content. The result showed that SVM is more accurate in detecting user offensiveness than naïve Bayes (NB). However, NB is faster than SVM. Chavan and Shylaja [19] proposed the use of SVM to build a classifier for the detection of cyberbullying in social networking sites. Data containing offensive words were extracted from social networking sites and utilized to build a cyberbullying SVM prediction model. The SVM classifier detected cyberbullying more accurately than LR did. Dadvar *et al.* [61] used SVM to build a gender specific cyberbullying prediction model. An SVM text classifier was created with gender specific characteristics.

The SVM cyberbullying prediction model enhanced the detection of cyberbullying in SM. Hee *et al.* [72] developed an SVM-based cyberbullying detection model to detect cyberbullying in a social network site. The SVM-based model was trained using data containing cyberbullying extracted from the social network site. The researchers found that that the SVM-based cyberbullying model effectively detected cyberbullying. Mangaonkar *et al.* [73] constructed an SVM-based cyberbullying detection model for YouTube. Data were collected from YouTube comments on videos posted on the site. The data were used to train SVM and construct a cyberbullying detection model, which was then used to detect cyberbullying. The results suggested that the SVM-based cyberbullying model is more reliable but not as accurate as rule-based Jrip. However, the SVM-based cyberbullying model is more accurate than NB and tree-based J48. Dinakar *et al.* [95] proposed the use of SVM for the detection of cyberbullying in Twitter. An SVM-based cyberbullying model was constructed from data extracted from Twitter. The SVM-based cyberbullying prediction model was applied to detect cyberbullying in Twitter. SVM detected cyberbullying better than NB- and LR-based cyberbullying detection models did.

2) NB ALGORITHM

NB was used to construct cyberbullying prediction models in [18], [38], [73], [74], and [95]. NB classifiers were constructed by applying Bayes' theorem between features. Bayesian learning is commonly used for text classification. This model assumes that the text is generated by a parametric model and utilizes training data to compute Bayes-optimal estimates of the model parameters. It categorizes generated test data with these approximations [112].

NB classifiers can deal with an arbitrary number of continuous or categorical independent features [106]. By using the assumption that the features are independent, a high-dimensional density estimation task is reduced to one-dimensional kernel density estimation [106].

The NB algorithm is a learning algorithm that is grounded on the use of Bayes theorem with strong (naive) independence assumptions. This method was discussed in detail in [113]. The NB algorithm is one of the most commonly used machine learning algorithms [114], and it has been constructed as a machine learning classifier in numerous social media based studies [115]–[117].

3) RANDOM FOREST

Random forest (RF) was used in the construction of cyberbullying prediction models in [72] and [86]. RF is a machine-learning model that combines decision trees and ensemble learning [118]. This model fits several classification trees to a dataset then combines the predictions from all the trees [119]. Therefore, RF consists of many trees that are used randomly to select feature variables for the classifier input. The construction of RF is achieved in the following simplified steps.

1. The number of examples (cases) in training data is set to N, and the number of attributes in the classifier is M.
2. A number of random decision tress is created by selecting attributes randomly. A training set is selected for each tree by choosing n times from all N existing instances. The rest of the instances in the training set are used to approximate the error of the tree by forecasting their classes.
3. For each tree's nodes, m random variables are selected on which to base the decision at that node. The finest split is computed using these m attributes in the training set. Each tree is completely built and is not pruned, as can be done in building a normal tree classifier.
4. A large number of trees are thus created. These decision trees vote for the most popular class. These processes are called RFs [118].

RF constructs a model that comprises a group of tree-structured classifiers, in which each tree votes for the most popular class [118]. The most highly voted class is the selected as the output.

4) DECISION TREE

Decision tree classifiers were used in construction of cyberbullying prediction models in [38] and [95]. Decision trees

are easy to understand and interpret; hence, the decision tree algorithm can be used to analyze data and build a graphic model for classification. The most commonly improved version of decision tree algorithms used for cyberbullying prediction is C.45 [38], [70], [95]. C4.5 can be explained as follows. Given N number of examples, C4.5 first produces an initial tree through the divide-and-conquer algorithm as follows [120]:

If all examples in N belong to the same class or N is small, the tree is a leaf labeled with the most frequent class in N. Otherwise, a test is selected based on, for example, the mostly used information gain test on a single attribute with two or more outputs. Considering that the test is the root of the tree creation partition of N into subsets $N_1, N_2, N_3 \ldots\ldots$ regarding the outputs for each example, the same procedure is applied recursively to each subset [120].

5) K-NEAREST NEIGHBOR

K-nearest neighbor (KNN) is a nonparametric technique that decides the KNNs of X_0 and uses a majority vote to calculate the class label of X_0. The KNN classifier often uses Euclidean distances as the distance metric [121]. To demonstrate a KNN classification, classifying new input posts (from a testing set) is considered by using a number of known manually labeled posts. The main task of KNN is to classify the unknown example based on a nominated number of its nearest neighbors, that is, to finalize the class of unknown examples as either a positive or negative class. KNN classifies the class of unknown examples by using majority votes for the nearest neighbors of the unknown classes. For example, if KNN is one nearest neighbor [estimating the class of an unknown example using the one nearest neighbor vote (k = 1)], then KNN will classify the class of the unknown example as positive (because the closest point is positive). For two nearest neighbors (estimating the class of an unknown example using the two nearest neighbor vote), KNN is unable to classify the class of the unknown example because the second closest point is negative (positive and negative votes are equal). For four nearest neighbors (estimating the class of an unknown example using the four nearest neighbor vote), KNN classifies the class of the unknown example as positive (because the three closest points are positive and only one vote is negative). The KNN algorithm is one of the simplest classification algorithms, but despite its simplicity, it can provide competitive results [122]. KNN was used in the construction of cyberbullying prediction models in [38].

6) LOGISTIC REGRESSION CLASSIFICATION

Logistic regression is one of the common techniques imported by machine learning from the statistics field. Logistic regression is an algorithm that builds a separating hyperplane between two datasets by means of the logistic function [123]. The logistic regression algorithm takes inputs (features) and generates a forecast according to the probability of the input being appropriate for a class. For example, if the probability is >0.5, the classification of the instance will be a positive class; otherwise, the prediction is for the other class (negative class) [124]. Logistic regression was used in the construction of cyberbullying prediction models in [19] and [73].

E. EVALUATION

The primary objective of constructing prediction models based on machine learning is to generalize more than the training dataset [79]. When a machine learning model is applied to a real example, it can perform well. Accordingly, the data are divided into two parts. The first part is the training data used to train machine learning algorithms. The second part is the testing data used to test machine learning algorithms. However, separately dividing data into training and testing is not widely employed [79], especially in applications in which deriving training and testing data are difficult. For example, in cyberbullying prediction, most state-of-art studies manually labeled data. Hence, creating labeled data is expensive. These issues can be reduced by cross validation, that is, randomly dividing the training data into 10 subsets for example, and this process is called 10-fold cross validation. Cross validation involves the following steps: keep a fold separate (the model does not see it) and train data on the model by using the remaining folds; test each learned classifier on the fold which it did not see; and average the results to see how well the particular parameter setting performs [79], [125].

F. EVALUATION METRICS

Researchers measure the effectiveness of a proposed model to determine how successfully the model can distinguish cyberbullying from non-cyberbullying by using various evaluation measures. Reviewing common evaluation metrics in the research community is important to understand the performance of conflicting models. The most commonly used metrics in evaluating cyberbullying classifiers for SM websites are as follows:

1) ACCURACY

It was used to evaluate cyberbullying prediction models in [62], [70], [73] and [95], and it is calculated as follows:

$$Accuracy = \frac{(tp + tn)}{(tp + fp + tn + fn)}. \tag{6}$$

2) PRECISION, RECALL, AND F-MEASURE

These were used to evaluate cyberbullying prediction models in [18], [61], [72], and [73]. They are calculated as follows:

$$Precision = \frac{tp}{(tp + fp)}, \tag{7}$$

$$Recall = \frac{tp}{(tp + fn)}, \tag{8}$$

$$F - Measure = \frac{2 \times precision \times recall}{recision + recall} \tag{9}$$

where tp means true positive, tn is true negative, fp denotes false positive, and fn is false negative.

3) AREA UNDER THE CURVE (AUC)

AUC offers a discriminatory rate of the classifier at various operating points [3], [19], [38]. The main benefit of using AUC as an evaluation metric is that AUC gives a more robust measurement than the accuracy metric in class-imbalance situations [19], [38].

III. ISSUES RELATED TO CONSTRUCTING CYBERBULLYING PREDICTION MODELS

In this section, the issues identified from the reviewed studies are discussed. The main issues related to cyberbullying definition, data collection feature engineering, and evaluation metric selection are identified and discussed in following subsections.

A. ISSUES RELATED TO CYBERBULLYING DEFINITION

Traditional bullying is generally defined as ''intentional behavior to harm another, repeatedly, where it is difficult for the victim to defend himself or herself'' [126]. By extending the definition of traditional bullying, cyberbullying has been defined [90] as ''an aggressive behavior that is achieved using electronic platforms by a group or an individual repeatedly and over time against a victim who cannot easily defend him or herself.'' Applying such a definition makes it difficult to classify manually labeled data (the instance in which machine learning algorithms learn from) and whether a post is cyberbullying or not. Two main issues make the above definition difficult to be applied in online environments [47], [127]. The first issue is how to measure ''repeatedly and over time aggressive behavior'' on SM, and the second one is how to measure power imbalance and ''a victim who cannot easily defend himself or herself'' on SM. These issues have been discussed by researchers to simplify the concept of cyberbullying in the online context. First, the concept of repetitive act in cyberbullying is not as straightforward as that in SM [47]. For example, SM websites can provide cyberbullies a medium to propagate cyberbullying posts for a large population. Consequently, a single act by one committer may become repetitive over time [47]. Second, power imbalance is presented in different forms in online communication. Researchers [127] have suggested that the content in online environments is difficult to eliminate or avoid, thus making a victim powerless.

These definitional aspects are under intense debate, but to simplify the definition of cyberbullying and make this definition applicable to a wide range of applications, the researchers in [53] and [72] defined cyberbullying as ''the use of electronic communication technologies to bully others.'' Proposing a simplified and clear definition of cyberbullying is a crucial step toward building machine learning models that can satisfy the definition criteria of cyberbullying engagement.

B. DATA COLLECTION

Many cyberbullying prediction studies extracted their datasets by using specific keywords or profile IDs. Nevertheless, by simply tracking posts that have particular keywords, these researches may have presented potential sampling bias [82], [128], limited the prediction to posts that contain the predefined keywords, and overlooked many other posts relevant to cyberbullying. Such data collection methods limit the prediction model of cyberbullying to specified keywords. The identification of keywords for extracting posts is also subject to the author's understanding of cyberbullying. An effective method should use a complete range of posts indicating cyberbullying to train the machine learning classifier and ensure the generalization capability of the cyberbullying prediction model [43]. An important objective of machine learning is to generalize and not to limit the examples in a training dataset [79]. Researchers should investigate whether the sampled data are extracted from data that effectively represents all possible activities on SM websites [128]. Extracting well-representative data from SM is the first step toward building effective machine learning prediction models. However, SM websites' public application program interface (API) only allows the extraction of a small sample of all relevant data and thus poses a potential for sampling bias [80]–[82]. For example, a previous study [128] discussed whether data extracted from Twitter's streaming API is a sufficient representation of the activities in the Twitter network as a whole; the author compared keyword (words, phrases, or hashtags), user ID, and geo-coded sampling. Twitter's streaming API returns a dataset with some bias when keyword or user ID sampling is used. By contrast, using geo-tagged filtering provides good data representation [128]. With these points in mind, researchers should ensure minimum bias as much as possible when they extract data to guarantee that the examples selected to be represented in training data are generalized and provide an effective model when applied to testing data. Bias in data collection can impose bias in the selected training dataset based on specific keywords or users, and such a bias consequently introduces overfitting issues that affect the capability of a machine learning model to make reliable predictions on untrained data.

C. FEATURE ENGINEERING

Features are vital components in improving the effectiveness of machine learning prediction models [79]. Most of the discussed studies attempted to provide effective machine learning solutions to cyberbullying on SM websites by providing significant features (Table 1). However, these studies overlooked other important features. For example, online cyberbullies may dynamically change the way they use words and acronyms. SM websites help create cyberbullying acronyms that have not been commonly used in committing traditional bullying or are beyond SM norms [129]. Recent survey response studies (questionnaire-based studies) have reported positive correlations between different variables, such as personality [93], [94] and sociability of a user in an online environment [130], and cyberbullying occurrences. The observations of these studies are important in understanding such behavior in online environments. However, these

observations are yet to be used as features with machine learning algorithms to provide significant models. These observations can be useful when transformed to a practical form (features) that can be employed to develop effective machine learning prediction models for cyberbullying on SM websites. The abundant information provided by SM websites should be utilized to convert observations into a set of features. For example, two studies [17], [61] attempted to improve machine learning classifier performance by including features, such as age and gender, that show improvement in classifier performance, but these features are extracted from direct user details mentioned in the online profiles of users. However, most studies found that only a few users provide complete details in their online profiles [131], [132]. These studies suggested the useful practice of utilizing words expressed in the content (posts) to identify user age and gender [131], [132]. Moreover, cyberbullying is related to the aggressive behavior of a user. A study demonstrated that aggression considerably predicts cyberbullying [92]. Similarly, cyberbullying behavior has a strong correlation with neuroticism [93], [94]. Therefore, predicting if a user has used words related to neuroticism may provide a useful feature to predict cyberbullying engagement.

A significant correlation has also been found between sociability of a user and cyberbullying engagement in online environments [130]. Users who are highly active in online environments are likely to engage in cyberbullying [133]. According to these observations, SM websites possess features that can be used as signals to measure the sociability of a user, such as number of friends, number of posts, URLs in posts, hashtags in posts, and number of users engaged in conversations (mentioned). The combination of these features with traditionally used ones, such as profanity features, can provide comprehensive discriminative features. The reviewed studies (Table 1) focused on using either a traditional feature model (e.g., bag-of-words) or information (e.g., age or gender) limited to user profile information (information written by users in their profile). Given that such information is limited, comprehensive features should be proposed to improve classifier performance.

Moreover, maintaining a precise and accurate process in constructing machine learning models from start (data collection) to end (evaluation metric selection) is important in ensuring that the proposed features hold significance in improving classifier performance. The following subsection analyzes other issues related to constructing effective machine learning models for cyberbullying prediction on SM websites.

D. MACHINE LEARNING ALGORITHM SELECTION

A machine learning algorithm is selected to be trained on proposed features. However, deciding which classifier performs best for a specific dataset is difficult. More than one machine learning algorithm should be tested to determine the best machine learning algorithm for a specific dataset. Three points may be used as guide to narrow the selection of machine learning algorithms to be tested. First, a specific literature on machine learning for cyberbullying detection is important in selecting a specified classifier. The preeminence of the classifier may be circumscribed to a given domain [134]. Therefore, general previous research and findings on machine learning can used as a guide to select a machine learning algorithm. Second, a literature review of text mining [135], [136] can be used as a guide. Third, a performance comparison of comprehensive datasets [137] can be used as basis to select machine learning algorithms. However, although these three points can be used as guide to narrow the selection of machine learning algorithms, researchers need to test many machine learning algorithms to identify the optimal classifier for an accurate predictive model.

E. IMBALANCED CLASS DISTRIBUTION

In many cases of real data, datasets naturally have imbalanced classes in which the normal class has a large number of instances and the abnormal class has a small number of instances in the dataset. Abnormal class instances are rare and difficult to be collected from real-world applications. Examples of imbalanced data applications are fraud detection, instruction detection, and medical diagnosis. Similarly, the number of cyberbullying posts is expected to be much less than the number of non-cyberbullying posts, and this assumption generates an imbalanced class distribution in the dataset in which the instances of non-cyberbullying contain much more posts than those of cyberbullying. Such cases can prevent the model from correctly classifying the examples. Many methods have been proposed to solve this issue, and examples include SMOTE [138] and weight adjustment (cost-sensitive technique) [139].

The SMOTE technique [138] is applied to avoid overfitting, which occurs when particular replicas of minority classes are added to the main dataset. A subdivision of data is reserved from the minority class as an example, and new synthetic similar classes are generated. These synthetic classes are then added to the original dataset. The created dataset is used to train the machine learning methods. The cost-sensitive technique is utilized to control the imbalance class [139]. It is based on creating is a cost matrix, which defines the costs experienced in false positives and false negatives.

F. EVALUATION METRIC SELECTION

Accuracy, precision, recall, and AUC are commonly used as evaluation metrics [19], [38]. Evaluation metric selection is important. The selection is based on the nature of manually labeled data. Selecting an inappropriate evaluation metric may result in better performance according to the selected evaluation metric. Then, the researcher may find the results to be significantly improved, although an investigation of how the machine learning model is evaluated may produce contradicting results and may not truly reflect the improvement of performance. For example, cyberbullying posts are commonly considered abnormal cases, whereas

non-cyberbullying posts are considered normal cases. The ratio between cyberbullying and non-cyberbullying is normally large. Generally, non-cyberbullying posts comprise a large portion. For example, 1000 posts are manually labeled as cyberbullying and non-cyberbullying. The non-cyberbullying posts are 900, and the remaining 100 posts are cyberbullying. If a machine learning classifier classifies all 1000 posts as non-cyberbullying and is unable to classify any posts (0) as cyberbullying, then this classifier is considered impractical. By contrast, if researchers use accuracy as the main evaluation metric, then the accuracy of this classifier calculated as mentioned in the accuracy equation will yield a high accuracy percentage.

In the example, the classifier fails to classify any cyberbullying posts but obtains a high accuracy percentage. Knowing the nature of manually labeled data is important in selecting an evaluation metric. In cases where data are imbalanced, researchers may need to select AUC as the main evaluation metric. In class-imbalance situations, AUC is more robust than other performance metrics [140]. Cyberbullying and non-cyberbullying data are commonly imbalanced datasets (non-cyberbullying posts outnumber the cyberbullying ones) that closely represent the real-life data that machine learning algorithms need to train on. Accordingly, the learning performance of these algorithms is independent of data skewness [73]. Special care should be taken in selecting the main evaluation metric to avoid uncertain results and appropriately evaluate the performance of machine learning algorithms.

IV. ISSUES AND CHALLENGES

This section presents the issues and challenges while guiding future researchers to explore the domain of sentiment analysis through leveraging machine learning algorithms and models for detecting cyberbullying through social media.

A. HUMAN DATA CHARACTERISTICS

Although SM big data provide insights into large human behavior data, in reality, the analysis of such big data remains subjective [141]. Building human prediction systems involves steps where subjectivity about human behavior does exist. For example, when creating a manually labeled dataset to train a machine learning algorithm to predict cyberbullying posts, human bias may exist based on how cyberbullying is being defined and the criteria used to categorize the text as cyberbullying text.

Moreover, subjectivity may exist during the creation of a set of features (learning factors) in the feature engineering process. For example, the pre-processing stage involves a ''data cleaning'' process wherein choices about what features will be counted, and which will be ignored are constructed. This process is inherently subjective [141].

Predicting human behavior is crucial but complex. To achieve an effective prediction of human behavior, the patterns that exist and are used for constructing the prediction model should also exist in the future input data. The patterns should clearly represent features that occur in current and future data to retain the context of the model. Given that big data are not generic and dynamic in nature, the context of these data is difficult to understand in terms of scale and even more difficult to maintain when data are reduced to fit into a machine learning model. Handling context of big data is challenging and has been presented as an important future direction [141].

Furthermore, human behavior is dynamic. Knowing when online users change the way of committing cyberbullying is an important component in updating the prediction model with such changes. Therefore, dynamically updating the prediction model is necessary to meet human behavioral changes [1].

B. CULTURE EFFECT

What was considered cyberbullying yesterday might not be considered cyberbullying today, and what was previously considered cyberbullying may not be considered cyberbullying now due to the introduction of OSNs. OSNs have a globalized culture. However, machine learning always learns from the examples provided. Consequently, designing different examples that represent a different culture remains to be defined, and robust work from different disciplines is required. For this purpose, cross disciplinary coordination is highly desirable.

C. LANGUAGE DYNAMICS

Language is quickly changing, particularly among the young generation. New slang is regularly integrated into the language culture. Therefore, researchers are invited to propose dynamic algorithms to detect new slang and abbreviations related to cyberbullying behavior on SM websites and keep updating the training processes of machine learning algorithms by using newly introduced words.

D. PREDICTION OF CYBERBULLYING SEVERITY

The level of cyberbullying severity should be determined. The effect of cyberbullying is proportional to its severity and spread. Predicting different levels of cyberbullying severity does not only require machine learning understanding but also a comprehensive investigation to define and categorize the level of cyberbullying severity from social and psychological perceptions. Efforts from different disciplines are required to define and identify the levels of severity then introduce related factors that can be converted into features to build multi-classifier machine learning for classifying cyberbullying severity into different levels as opposed to a binary classifier that only detects whether an instance is cyberbullying or not.

E. UNSUPERVISED MACHINE LEARNING

Human learning is essentially unsupervised. The structure of the world was discovered by observing it and not by being told the name of every objective. Nevertheless, unsupervised machine learning has been overshadowed by the success

of supervised learning [142]. This gap in literature may be caused by the fact that nearly all current studies rely on manually labeled data as the input to supervised algorithms for classifying classes. Thus, finding patterns between two classes by using unsupervised grouping remains difficult. Intensive research is required to develop unsupervised algorithms that can detect effective patterns from data. Traditional machine learning algorithms lack the capability to handle cyberbullying big data.

Deep learning has recently attracted the attention of many researchers in different fields. Natural language understanding is a new area in which deep learning is poised to make a large effect over the next few years [142].

The traditional machine learning algorithms pointed out in this survey lacks the capability to process big data in a standalone format. Big data have rendered traditional machine learning algorithms impotent. Cyberbullying big data generated from SM require advanced technology for the processing of the generated data to gain insights and help in making intelligent decisions.

Big data are generated at a very high velocity, variety, volume, verdict, value, veracity, complexity, etc. Researchers need to leverage various deep learning techniques for processing social media big data for cyberbullying behaviors. The deep learning techniques and architectures with a potential to explore the cyberbullying big data generated from SM can include generative adversarial network, deep belief network, convolutional neural network, stacked autoencoder, deep echo state network, and deep recurrent neural network. These deep learning architectures remain unexplored in cyberbullying detection in SM.

V. CONCLUSIONS AND FUTURE DIRECTIONS

This study reviewed existing literature to detect aggressive behavior on SM websites by using machine learning approaches. We specifically reviewed four aspects of detecting cyberbullying messages by using machine learning approaches, namely, data collection, feature engineering, construction of cyberbullying detection model, and evaluation of constructed cyberbullying detection models. Several types of discriminative features that were used to detect cyberbullying in online social networking sites were also summarized. In addition, the most effective supervised machine learning classifiers for classifying cyberbullying messages in online social networking sites were identified.

One of the main contributions of current paper is the definition of evaluation metrics to successfully identify the significant parameter so the various machine learning algorithms can be evaluated against each other. Most importantly we summarized and identified the important factors for detecting cyberbullying through machine learning techniques specially supervised learning. For this purpose, we have used accuracy, precision recall and f-measure which gives us the area under the curve function for modeling the behaviors in cyberbullying. Finally, the main issues and open research challenges were described and discussed.

Considerable research effort is required to construct highly effective and accurate cyberbullying detection models. We believe that the current study will provide crucial details on and new directions in the field of detecting aggressive human behavior, including cyberbullying detection in online social networking sites.

REFERENCES

[1] V. Subrahmanian and S. Kumar, "Predicting human behavior: The next frontiers," *Science*, vol. 355, no. 6324, p. 489, 2017.
[2] H. Lauw, J. C. Shafer, R. Agrawal, and A. Ntoulas, "Homophily in the digital world: A LiveJournal case study," *IEEE Internet Comput.*, vol. 14, no. 2, pp. 15–23, Mar./Apr. 2010.
[3] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Comput. Hum. Behav.*, vol. 63, pp. 433–443, Oct. 2016.
[4] L. Phillips, C. Dowling, K. Shaffer, N. Hodas, and S. Volkova, "Using social media to predict the future: A systematic literature review," 2017, *arXiv:1706.06134*. [Online]. Available: https://arxiv.org/abs/1706.06134
[5] H. Quan, J. Wu, and Y. Shi, "Online social networks & social network services: A technical survey," in *Pervasive Communication Handbook*. Boca Raton, FL, USA: CRC Press, 2011, p. 4.
[6] J. K. Peterson and J. Densley, "Is social media a gang? Toward a selection, facilitation, or enhancement explanation of cyber violence," *Aggression Violent Behav.*, 2016.
[7] BBC. (2012). *Huge Rise in Social Media*. [Online]. Available: http://www.bbc.com/news/uk-20851797
[8] P. A. Watters and N. Phair, "Detecting illicit drugs on social media using automated social media intelligence analysis (ASMIA)," in *Cyberspace Safety and Security*. Berlin, Germany: Springer, 2012, pp. 66–76.
[9] M. Fire, R. Goldschmidt, and Y. Elovici, "Online social networks: Threats and solutions," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2019–2036, 4th Quart., 2014.
[10] N. M. Shekokar and K. B. Kansara, "Security against sybil attack in social network," in *Proc. Int. Conf. Inf. Commun. Embedded Syst. (ICICES)*, 2016, pp. 1–5.
[11] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, "Detecting and tracking political abuse in social media," in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011, pp. 297–304.
[12] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, "PhishAri: Automatic realtime phishing detection on Twitter," in *Proc. eCrime Res. Summit (eCrime)*, Oct. 2012, pp. 1–12.
[13] S. Yardi *et al.*, "Detecting spam in a Twitter network," *First Monday*, Jan. 2009. [Online]. Available: https://firstmonday.org/article/view/2793/2431
[14] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, "Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on twitter," in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 71–80.
[15] G. R. S. Weir, F. Toolan, and D. Smeed, "The threats of social networking: Old wine in new bottles?" *Inf. Secur. Tech. Rep.*, vol. 16, no. 2, pp. 38–43, 2011.
[16] M. J. Magro, "A review of social media use in e-government," *Administ. Sci.*, vol. 2, no. 2, pp. 148–161, 2012.
[17] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in *Advances in Information Retrieval*. Berlin, Germany: Springer, 2013, pp. 693–696.
[18] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *Proc. Int. Conf. Privacy, Secur., Risk Trust (PASSAT)*, Sep. 2012, pp. 71–80.
[19] V. S. Chavan and S. S. Shylaja, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," in *Proc. Int. Conf. Adv. Comput., Commun. Inform. (ICACCI)*, Aug. 2015, pp. 2354–2358.
[20] W. Dong, S. S. Liao, Y. Xu, and X. Feng, "Leading effect of social media for financial fraud disclosure: A text mining based analytics," in *Proc. AMCIS*, San Diego, CA, USA, 2016.
[21] M. S. Rahman, T.-K. Huang, H. V. Madhyastha, and M. Faloutsos, "FRAppE: Detecting malicious Facebook applications," in *Proc. 8th Int. Conf. Emerg. Netw. Exp. Technol.*, 2012, pp. 313–324.

[22] S. Abu-Nimeh, T. Chen, and O. Alzubi, "Malicious and spam posts in online social networks," *Computer*, vol. 44, no. 9, pp. 23–28, Sep. 2011.
[23] B. Doerr, M. Fouz, and T. Friedrich, "Why rumors spread so quickly in social networks," *Commun. ACM*, vol. 55, no. 6, pp. 70–75, Jun. 2012.
[24] J. W. Patchin and S. Hinduja, *Words Wound: Delete Cyberbullying and Make Kindness Go Viral*. Golden Valley, MN, USA: Free Spirit Publishing, 2013.
[25] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," in *Proc. 9th Int. AAAI Conf. Web Social Media*, Apr. 2015.
[26] S. Liu, J. Zhang, and Y. Xiang, "Statistical detection of online drifting Twitter spam: Invited paper," in *Proc. 11th ACM Asia Conf. Comput. Commun. Secur.*, 2016, pp. 1–10.
[27] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," *Inf. Sci.*, vol. 260, pp. 64–73, Mar. 2014.
[28] M. Jiang, S. Kumar, V. S. Subrahmanian, and C. Faloutsos, "KDD 2017 tutorial: Data-driven approaches towards malicious behavior modeling," *Dimensions*, vol. 19, p. 42, 2017.
[29] S. Y. Jeong, Y. S. Koh, and G. Dobbie, "Phishing detection on Twitter streams," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2016, pp. 141–153.
[30] I. Frommholz, H. M. Al-Khateeb, M. Potthast, Z. Ghasem, M. Shukla, and E. Short, "On textual analysis and machine learning for cyberstalking detection," *Datenbank-Spektrum*, vol. 16, no. 2, pp. 127–135, 2016.
[31] M. McCord and M. Chuah, "Spam detection on Twitter using traditional classifiers," in *Autonomic and Trusted Computing*. Berlin, Germany: Springer, 2011, pp. 175–186.
[32] X. Chen, R. Chandramouli, and K. P. Subbalakshmi, "Scam detection in Twitter," in *Data Mining for Service*. Berlin, Germany: Springer, 2014, pp. 133–150.
[33] A. H. Wang, "Detecting spam bots in online social networking sites: A machine learning approach," in *Data and Applications Security and Privacy XXIV*. Berlin, Germany: Springer, 2010, pp. 335–342.
[34] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," *Neurocomputing*, vol. 159, pp. 27–34, Jul. 2015.
[35] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, p. 18, 2012.
[36] M. Dadvar and F. De Jong, "Cyberbullying detection: A step toward a safer Internet yard," in *Proc. 21st Int. Conf. Companion World Wide Web*, 2012, pp. 121–126.
[37] S. O. Sood, J. Antin, and E. Churchill, "Using crowdsourcing to improve profanity detection," in *Proc. AAAI Spring Symp., Wisdom Crowd*, 2012, pp. 69–74.
[38] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying," in *Proc. Int. Joint Conf. SOCO-CISIS-ICEUTE*. Cham, Switzerland: Springer, 2014, pp. 419–428.
[39] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis,"in *Proc. 3rd Int. Workshop Socially-Aware Multimedia*, 2014, pp. 3–6.
[40] R. M. Kowalski, *Cyberbullying: Bullying in the Digital Age*. Hoboken, NJ, USA: Wiley, 2012.
[41] T. Nakano, T. Suda, Y. Okaie, and M. J. Moore, "Analysis of cyber aggression and cyber-bullying in social networking," in *Proc. IEEE 10th Int. Conf. Semantic Comput. (ICSC)*, Feb. 2016, pp. 337–341.
[42] G. S. O'Keeffe and K. Clarke-Pearson, "The impact of social media on children, adolescents, and families," *Pediatrics*, vol. 127, no. 4, pp. 800–804, 2011.
[43] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2012, pp. 656–666.
[44] R. M. Kowalski and S. P. Limber, "Psychological, physical, and academic correlates of cyberbullying and traditional bullying," *J. Adolescent Health*, vol. 53, no. 1, pp. S13–S20, 2013.
[45] H. Sampasa-Kanyinga, P. Roumeliotis, and H. Xu, "Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among Canadian schoolchildren," *PLoS ONE*, vol. 9, no. 7, 2014, Art. no. e102145.
[46] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Arch. Suicide Res.*, vol. 14, no. 3, pp. 206–221, 2010.
[47] R. Slonje, P. K. Smith, and A. Frisén, "The nature of cyberbullying, and strategies for prevention," *Comput. Hum. Behav.*, vol. 29, no. 1, pp. 26–32, 2013.
[48] E. Whittaker and R. M. Kowalski, "Cyberbullying via social media," *J. School Violence*, vol. 14, no. 1, pp. 11–29, 2015.
[49] F. Sticca and S. Perren, "Is cyberbullying worse than traditional bullying? Examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying," *J. Youth Adolescence*, vol. 42, no. 5, pp. 739–750, 2013.
[50] S. Wen, J. Jiang, Y. Xiang, S. Yu, W. Zhou, and W. Jia, "To shut them up or to clarify: Restraining the spread of rumors in online social networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 12, pp. 3306–3316, Dec. 2014.
[51] K. Van Royen, K. Poels, W. Daelemans, and H. Vandebosch, "Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability," *Telematics Inform.*, vol. 32, no. 1, pp. 89–97, 2015.
[52] K. Van Royen, K. Poels, and H. Vandebosch, "Harmonizing freedom and protection: Adolescents' voices on automatic monitoring of social networking sites," *Children Youth Services Rev.*, vol. 64, pp. 35–41, May 2016.
[53] R. M. Kowalski, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth," *Psychol. Bull.*, vol. 140, no. 4, pp. 1073–1137, 2014.
[54] Q. Li, "New bottle but old wine: A research of cyberbullying in schools," *Comput. Hum. Behav.*, vol. 23, no. 4, pp. 1777–1791, 2007.
[55] R. S. Tokunaga, "Following you home from school: A critical review and synthesis of research on cyberbullying victimization," *Comput. Hum. Behav.*, vol. 26, no. 3, pp. 277–287, May 2010.
[56] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2014.
[57] Y. Liu, J. Yang, Y. Huang, L. Xu, S. Li, and M. Qi, "MapReduce based parallel neural networks in enabling large scale machine learning," *Comput. Intell. Neurosci.*, vol. 2015, p. 1, Jan. 2015.
[58] C. Wu, R. Buyya, and K. Ramamohanarao, "Big data analytics = machine learning + cloud computing," 2016, *arXiv:1601.03115*. [Online]. Available: https://arxiv.org/abs/1601.03115
[59] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "A survey on deep learning for big data," *Inf. Fusion*, vol. 42, pp. 146–157, Jul. 2018.
[60] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, 2015.
[61] M. Dadvar, F. D. Jong, R. Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information," in *Proc. 25th Dutch-Belgian Inf. Retr. Workshop*, 2012, pp. 1–3.
[62] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Detection of cyberbullying incidents on the instagram social network," 2015, *arXiv:1503.03909*. [Online]. Available: https://arxiv.org/abs/1503.03909
[63] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, pp. 1289–1305, Mar. 2003.
[64] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proc. ACL*, 2002, pp. 417–424.
[65] R. M. Tong, "An operational system for detecting and tracking opinions in on-line discussion," in *Proc. Notes ACM SIGIR Workshop Oper. Text Classification*, 2001.
[66] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proc. 8th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 1997, pp. 174–181.
[67] S. Nadali, M. A. A. Murad, N. M. Sharef, A. Mustapha, and S. Shojaee, "A review of cyberbullying detection: An overview," in *Proc. 13th Int. Conf. Intell. Syst. Design Appl. (ISDA)*, Dec. 2013, pp. 325–330.
[68] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting cyberbullying: Query terms and techniques," in *Proc. 5th Annu. ACM Web Sci. Conf.*, 2013, pp. 195–204.
[69] H. Chen, S. Mckeever, and S. J. Delany, "Harnessing the power of text mining for the detection of abusive content in social media," in *Advances in Computational Intelligence Systems*. Cham, Switzerland: Springer, 2017, pp. 187–205.
[70] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops (ICMLA)*, Dec. 2011, pp. 241–244.

[71] V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection," *Commun. Inf. Sci. Manage. Eng.*, vol. 3, no. 5, p. 238, 2013.
[72] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, and V. Hoste, "Detection and fine-grained classification of cyberbullying events," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process. (RANLP)*, 2015, pp. 672–680.
[73] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," in *Proc. IEEE Int. Conf. Electro/Inf. Technol. (EIT)*, Dekalb, IL, USA, May 2015, pp. 611–616.
[74] H. Sanchez and S. Kumar, "Twitter bullying detection," Tech. Rep. UCSC ISM245, 2011.
[75] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Mach. Learn.*, vol. 50, nos. 1–2, pp. 5–43, 2003.
[76] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Rev. Genet.*, vol. 16, no. 6, pp. 321–332, 2015.
[77] P. Langley and H. A. Simon, "Applications of machine learning and rule induction," *Commun. ACM*, vol. 38, no. 11, pp. 54–64, 1995.
[78] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452–459, May 2015.
[79] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.
[80] Y. Liu, C. Kliman-Silver, and A. Mislove, "The tweets they are a-changin': Evolution of twitter users and behavior," in *Proc. Int. AAAI Conf. Weblogs Social Media (ICWSM)*, 2014, pp. 305–314.
[81] S. González-Bailón, N. Wang, A. Rivero, J. Borge-Holthoefer, and Y. Moreno, "Assessing the bias in samples of large online networks," *Social Netw.*, vol. 38, pp. 16–27, Jul. 2014.
[82] T. Cheng and T. Wicks, "Event detection using Twitter: A spatio-temporal approach," *PLoS ONE*, vol. 9, no. 6, p. e97807, 2014.
[83] A. Bellmore, A. J. Calvin, J.-M. Xu, and X. Zhu, "The five W's of 'bullying' on Twitter: Who, what, why, where, and when," *Comput. Hum. Behav.*, vol. 44, pp. 305–314, Mar. 2015.
[84] H. Margono, X. Yi, and G. K. Raikundalia, "Mining Indonesian cyber bullying patterns in social networks," in *Proc. 37th Australas. Comput. Sci. Conf.*, vol. 147, Australian Computer Society, 2014, pp. 115–124.
[85] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proc. 17th Int. Conf. Distrib. Comput. Netw.*, 2016, Art. no. 43.
[86] Á. García-Recuero, "Discouraging abusive behavior in privacy-preserving online social networking applications," in *Proc. 25th Int. Conf. Companion World Wide Web*, International World Wide Web Conferences Steering Committee, 2016, pp. 305–309.
[87] Y. Anzai, *Pattern Recognition and Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2012.
[88] E. Calvete, I. Orue, A. Estévez, L. Villardón, and P. Padilla, "Cyberbullying in adolescents: Modalities and aggressors' profile," *Comput. Hum. Behav.*, vol. 26, no. 5, pp. 1128–1135, 2010.
[89] H. Vandebosch and K. Van Cleemput, "Cyberbullying among youngsters: Profiles of bullies and victims," *New Media Soc.*, vol. 11, no. 8, pp. 1349–1371, 2009.
[90] R. Slonje and P. K. Smith, "Cyberbullying: Another main type of bullying?" *Scand. J. Psychol.*, vol. 49, no. 2, pp. 147–154, 2008.
[91] K. R. Williams and N. G. Guerra, "Prevalence and predictors of Internet bullying," *J. Adolescent Health*, vol. 41, no. 6, pp. S14–S21, 2007.
[92] O. T. Arıcak, "Psychiatric symptomatology as a predictor of cyberbullying among University Students," *Eurasian J. Educ. Res.*, vol. 34, no. 1, p. 169, 2009.
[93] I. Connolly and M. O'Moore, "Personality and family relations of children who bully," *Personality Individual Differences*, vol. 35, no. 3, pp. 559–567, 2003.
[94] L. Corcoran, I. Connolly, and M. O'Moore, "Cyberbullying in Irish schools: An investigation of personality and self-concept," *Irish J. Psychol.*, vol. 33, no. 4, pp. 153–165, 2012.
[95] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011, pp. 11–17.
[96] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on Web 2.0," in *Proc. Content Anal. Web*, 2009, pp. 1–7.
[97] M. Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, and K. Araki, "Machine learning and affect analysis against cyber-bullying," in *Proc. 36th AISB*, 2010, pp. 7–16.
[98] E. Raisi and B. Huang, "Cyberbullying identification using participant-vocabulary consistency," 2016, *arXiv:1606.08084*. [Online]. Available: https://arxiv.org/abs/1606.08084
[99] A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin, "Identification and characterization of cyberbullying dynamics in an online social network," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, 2015, pp. 280–285.
[100] R. M. Gray, "Entropy and information," in *Entropy and Information Theory*. New York, NY, USA: Springer, 1990, pp. 21–55.
[101] I. Qabajeh and F. Thabtah, "An experimental study for assessing email classification attributes using feature selection methods," in *Proc. 3rd Int. Conf. Adv. Comput. Sci. Appl. Technol. (ACSAT)*, Dec. 2014, pp. 125–132.
[102] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*. Berlin, Germany: Springer, 2009, pp. 1–4.
[103] M. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 359–366.
[104] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 80–89, 2004.
[105] D. H. Wolper and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997.
[106] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016.
[107] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, 1998.
[108] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Dept. Comput. Sci., Nat. Taiwan Univ., Tech. Rep., 2003. [Online]. Available: http://www.csie. ntu.edu.tw/*cjlin/papers/guide/guide.pdf
[109] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Nov. 2001.
[110] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2013.
[111] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learn. Theory*, 1992, pp. 144–152.
[112] A. McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," in *Proc. AAAI Workshop Learn. Text Categorization*, 1998, pp. 1–8.
[113] H. Zhang, "The optimality of naive Bayes," Tech. Rep., 2004.
[114] H. Zhang, "The optimality of naive Bayes," in *Proc. IAAA*, vol. 1, no. 2, 2004, p. 3.
[115] N. Bora, V. Zaytsev, Y.-H. Chang, and R. Maheswaran "Gang networks, neighborhoods and holidays: Spatiotemporal patterns in social media," in *Proc. Int. Conf. Social Comput. (SocialCom)*, Sep. 2013, pp. 93–101.
[116] A. H. Wang, "Don't follow me: Spam detection in Twitter," in *Proc. Int. Conf. Secur. Cryptogr. (SECRYPT)*, Jul. 2010, pp. 1–10.
[117] D. M. Freeman, "Using naive bayes to detect spammy names in social networks," in *Proc. ACM Workshop Artif. Intell. Secur.*, 2013, pp. 3–12.
[118] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[119] D. R. Cutler, D. R. Cutler, T. C. Edwards, Jr., K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler, "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007.
[120] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.
[121] P. Soucy and G. W. Mineau, "A simple KNN algorithm for text categorization," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov./Dec. 2001, pp. 647–648.
[122] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, "Efficient kNN classification algorithm for big data," *Neurocomputing*, vol. 195, pp. 143–148, Jun. 2016.
[123] S. Dreiseitl, L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt, and M. Binder, "A comparison of machine learning methods for the diagnosis of pigmented skin lesions," *J. Biomed. Informat.*, vol. 34, no. 1, pp. 28–36, 2001.
[124] D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, vol. 398. Hoboken, NJ, USA: Wiley, 2013.
[125] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. IJCAI*, 1995, pp. 1137–1145.

[126] P. K. Smith, *The Nature of School Bullying: A Cross-National Perspective*. London, U.K.: Psychology Press, 1999.
[127] J. J. Dooley, J. Pyżalski, and D. Cross, "Cyberbullying versus face-to-face bullying: A theoretical and conceptual review," *Zeitschrift Psychologie/J. Psychol.*, vol. 217, no. 4, pp. 182–188, 2009.
[128] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? Comparing data from twitter's streaming API with Twitter's firehose," 2013, *arXiv:1306.5204*. [Online]. Available: https://arxiv.org/abs/1306.5204
[129] *From IHML (I Hate My Life) to MOS (Mum Over Shoulder): Why This Guide to Cyber-Bullying Slang May Save Your Child's Life*, Dailymail, USA, 2014. [Online]. Available: https://www.dailymail.co.uk/news/article-2673678/Why-guide-cyber-bullying-slang-save-childs-life-From-IHML-I-hate-life-Mos-mum-shoulder.html
[130] J. N. Navarro and J. L. Jasinski, "Going Cyber: Using routine activities theory to predict cyberbullying experiences," *Sociol. Spectr.*, vol. 32, no. 1, pp. 81–94, 2012.
[131] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting age and gender in online social networks," in *Proc. 3rd Int. Workshop Search Mining User-Generated Contents*, 2011, pp. 37–44.
[132] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, "'How old do you think I Am?' A study of language and age in Twitter," in *Proc. 7th Int. AAAI Conf. Weblogs Social Media*, AAAI Press, 2013, pp. 439–448.
[133] V. Balakrishnan, "Cyberbullying among young adults in Malaysia: The roles of gender, age and Internet frequency," *Comput. Hum. Behav.*, vol. 46, pp. 149–157, May 2015.
[134] N. Maciá, E. Bernadó-Mansilla, A. Orriols-Puig, and T. K. Ho, "Learner excellence biased by data set selection: A case for data characterisation and artificial data sets," *Pattern Recognit.*, vol. 46, no. 3, pp. 1054–1066, Mar. 2013.
[135] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surveys*, vol. 34, no. 1, pp. 1–47, 2002.
[136] V. Korde and C. N. Mahender, "Text classification and classifiers: A survey," *Int. J. Artif. Intell. Appl.*, vol. 3, no. 2, p. 85, 2012.
[137] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *J. Mach. Learn. Res*, vol. 15, no. 1, pp. 3133–3181, 2014.
[138] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.
[139] X.-Y. Liu and Z.-H. Zhou, "The influence of class imbalance on cost-sensitive learning: An empirical study," in *Proc. 6th Int. Conf. Data Mining (ICDM)*, Dec. 2006, pp. 970–974.
[140] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[141] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Inf., Commun. Soc.*, vol. 15, no. 5, pp. 662–679, 2012.
[142] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

MOHAMMED ALI AL-GARADI received the Ph.D. degree from the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. He has published several articles in academic journals indexed in well-reputed databases such as ISI and Scopus. His research interests include cybersecurity, online social networking, machine learning text mining, deep learning, and the IoT.

MOHAMMAD RASHID HUSSAIN received the Ph.D. degree in information technology from Babasaheb Bhimrao Ambedkar Bihar University, India. He is currently an Assistant Professor with the College of Computer Science, King Khalid University, Abha, Saudi Arabia. His research interests include educational development & review, educational data mining, cloud intelligence, and mobile cloud computing and optimizations.

NAWSHER KHAN received the Ph.D. degree from University Malaysia Pahang (UMP), Malaysia, in 2013. He was a Postdoctoral Research Fellow with the University of Malaya (UM), Malaysia, in 2014. In 2005, he was appointed in National Database and Registration Authority (NADRA) under the Interior Ministry of Pakistan. In 2008, he has worked in National Highways Authority (NHA). He has served at Abdul Wali Khan University Mardan, Pakistan, as an Assistant Professor for 3 years, from 2014 to 2017. He is currently serving as an Associate Professor and the Director of Research Center, College of Computer Science, King Khalid University, Abha, Saudi Arabia. He has published more than 50 articles in various International Journals and Conference proceedings. His research interests include big data, cloud computing, data management, distributed systems, scheduling, replication, and sensor networks.

GHULAM MURTAZA is currently pursuing the Ph.D. degree with the Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. He is also an Assistant Professor with Sukkur IBA University, Sukkur, Pakistan. He is currently on study leave to pursue his Ph.D. He has published several articles in well-reputed databases. His research interests include machine learning, deep learning, digital image processing, big data, and information retrieval.

HENRY FRIDAY NWEKE received the B.Sc. degree in computer science from Ebonyi State University, Nigeria, and the M.Sc. degree in computer science from the University of Bedfordshire, U.K. He is currently pursuing the Ph.D. degree with the Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. His research interests include machine learning, deep learning, biomedical sensor analytics, human activity recognition, multi-sensor fusion, cloud computing, wireless sensor technologies, and emerging technology.

IHSAN ALI received the M.Sc. degree from Hazara University Manshera, Pakistan, in 2005, and the M.S. degree in computer system engineering from the GIK Institute, in 2008. He is currently pursuing the Ph.D. degree with the Faculty of Computer Science and Information Technology, University of Malaya. He has more than five years teaching and research experience in different country, including Saudi Arabia, USA, Pakistan, and Malaysia. He has served as a Technical Program Committee Member for the IWCMC 2017, AINIS 2017, Future 5V 2017, and also an Organizer of Special session on fog computing in Future 5V 2017. He has published more than 30 papers in the international journals and conferences. His research interests include wireless sensor networks, underwater sensor networks, sensor cloud, fog computing, and the IoT. He is also an Active Member of the IEEE, ACM, the International Association of Engineers (IAENG), and the Institute of Research Engineers and Doctors (the IRED). He is also a Reviewer of *Computers & Electrical Engineering*, *KSII Transactions on Internet and Information Systems*, *Mobile Networks and Applications*, the *International Journal of Distributed Sensor Networks*, the IEEE Transactions on Intelligent Transportation Systems, *Computer Networks*, IEEE Access, FGCS, and the *IEEE Communication Magazine*.

GHULAM MUJTABA received the master's degree in computer science from FAST National University, Karachi, Pakistan, and the Ph.D. degree from the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. He has received the gold medal for the masterâĂŹs degree. He has been an Assistant Professor with Sukkur IBA University, Sukkur, Pakistan, since 2006. He has vast experience in teaching and research. Before he joined Sukkur IBA University, he was with a well-known software house in Karachi for four years. He has also published several articles in academic journals indexed in well-reputed databases such as ISI and Scopus. His research interests include machine learning, online social networking, text mining, deep learning, and information retrieval.

HARUNA CHIROMA received the B.Tech. degree from Abubakar Tafawa Balewa University, the M.Sc. degree from Bayero University Kano, and the Ph.D. degree from the University of Malaya, all in computer science, and the Post-Graduate Diploma in education from Usman Danfodio University. He was a Visiting Senior Lecturer with Abubakar Tafawa Balewa University, Bauchi, and a Lecturer with the Federal College of Education, Gombe. As a teacher, he has developed interest in advanced learning technology. He has published more than 100 academic articles in international refereed ISI WoS Journals, Edited Books, Conference Proceedings, and Local Journals. He participated in the 2017 and 2018 QS world universities ranking evaluation of world universities research strengths in computer science. His research interests include deep learning, nature-inspired algorithms for machine learning, with special focus on their applications to internet of vehicles, autonomous vehicles, the Internet of Things, big data analytics, edge computing, cybersecurity, fog computing, and cloud computing. He has served in various capacities in more than 20 international conferences across the world. He is a member of the ACM, INNS, NCS, IAENG, and the Association for Computing Machinery (ACM). He is an Associate Editor of the IEEE Access and ISI WoS indexed Journal. He is a Leading Volume Editor of an edited book *Advances on Computational Intelligence in Energy-the Applications of Nature-Inspired & Metaheuristic Algorithms in Energy* (Heidelberg, Berlin: Springer), renowned series of Lecture Notes in Energy. He is a Reviewer for 15 ISI WoS indexed journals, such as *Applied Energy Q1* (Elsevier), *Applied Soft Computing Q1* (Elsevier), *Knowledge Based System Q1* (Elsevier), *Energy and Building Q1* (Elsevier), *Neural Computing and Applications Q1* (Springer), the *Journal of the Operational Research Society* (Springer), and *PLOS One*.

HASAN ALI KHATTAK received the B.Sc. degree in computer science from the University of Peshawar, Peshawar, Pakistan, in 2006, the master's degree in information engineering from the Politecnico di Torino, Torino, Italy, in 2011, and the Ph.D. degree in electrical and computer engineering from the Politecnico di Bari, Bari, Italy, in 2015. He has been serving as an Assistant Professor of computer science, since 2016. He is involved in a number of funded research projects in the Internet of Things, semantic web, and fog computing while exploring Ontologies, Web Technologies using Contiki OS, NS 2/3, and Omnet++ frameworks. His current research interests include distributed systems, web of things and vehicular ad hoc networks, and data and social engineering for smart cities.

ABDULLAH GANI received the Diploma degree in computer science from ITM, the B.Phil. and M.Sc. degrees in information management from the University of Hull, U.K., and the Ph.D. degree in computer science from the University of Sheffield, U.K. He acquired the Teaching Certificate from Kinta Teaching College, Ipoh. He has vast teaching experience due to having worked in a number of educational institutions locally and abroad–schools, the Malay Women Teaching College, Melaka, Ministry of Education, the Rotterham College of Technology and Art, Rotterham, U.K., and the University of Sheffield. He is currently a Professor with the Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. Since then, more than 150 academic papers have been published in proceedings and respectable journals internationally within top 10% ranking. He received a very good number of citation in Web of Science and Scopus databases and actively supervises. His interest in research kicked off, in 1983 when he was chosen to attend the 3 month Scientific Research Course in RECSAM by the Ministry of Education, Malaysia.

C. TOOLS USED

- Programming Platform:
    1. Django Framework
    2. Visual Studio Code
- Frontend Languages:
    1. HTML
    2. CSS
    3. JavaScript
- Backend language:
    1. Python

D. PARTICIPATION CERTIFICATE

Gokhale Education Society's
R. H. Sapat College of Engineering, Management Studies and Research Nashik-422005
Department of Electronics and Telecommunication Engineering

CERTIFICATE

OF APPRECIATION

THIS CERTIFICATE IS AWARDED TO

Miss. Rohini Jadhav

for participating in Project competition of G-ESTRONICA Organized by department of Electronics & Telecommunication Engineering in association with Electronics & Telecommunication Engineering Student's Technical Alliance (ESTA) during 20 May 2022-23

(Dr. S. P. Agnihotri)	(Dr. P. C. Kulkarni)	(Prof. P. M. Deshpande)
Head, E&TC	Principal	Director (P)